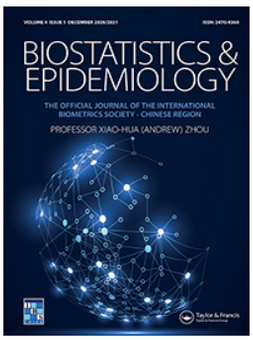


Taylor & Francis
Taylor & Francis Group

Biostatistics & Epidemiology

ISSN: (Print) (Online) Journal homepage: https://www.tandfonline.com/loi/tbep20

Intervention differential effects and regression to the mean in studies where sample selection is based on the initial value of the outcome variable: an evaluation of methods illustrated in weight-management studies

Lucy Beggs, Rebecca Briscoe, Claire Griffiths, George T. H. Ellison & Mark S. Gilthorpe

To cite this article: Lucy Beggs, Rebecca Briscoe, Claire Griffiths, George T. H. Ellison & Mark S. Gilthorpe (2020) Intervention differential effects and regression to the mean in studies where sample selection is based on the initial value of the outcome variable: an evaluation of methods illustrated in weight-management studies, Biostatistics & Epidemiology, 4:1, 172-188, DOI: 10.1080/24709360.2020.1719690

To link to this article: https://doi.org/10.1080/24709360.2020.1719690

Published online: 22 Mar 2020.

BIOSTATISTICS & EPIDEMIOLOGY
2020, VOL. 4, NO. 1, 172–188
https://doi.org/10.1080/24709360.2020.1719690

Taylor & Francis
Taylor & Francis Group

Intervention differential effects and regression to the mean in studies where sample selection is based on the initial value of the outcome variable: an evaluation of methods illustrated in weight-management studies

Lucy Beggs [a,b,c], Rebecca Briscoe [a,b,d], Claire Griffiths [a,e], George T. H. Ellison [a,b] and Mark S. Gilthorpe [a,b,f]

[a]Leeds Institute for Data Analytics, University of Leeds, Leeds, United Kingdom; [b]School of Medicine, University of Leeds, Leeds, United Kingdom; [c]National Institute for Health and Care Excellence, Manchester, United Kingdom; [d]Leeds Teaching Hospitals Trust, Leeds, United Kingdom; [e]Sports Science, Leeds Becket University, Leeds, United Kingdom; [f]The Alan Turing Institute, London, United Kingdom

ABSTRACT

Background: Intervention differential effects (IDEs) occur where changes in an outcome depend upon the initial values of that outcome. Although methods to identify IDEs are well documented, there remains a lack of understanding about the circumstances under which these methods are robust. One context that has not been explored is the identification of intervention differential effect in studies where sample selection is based on the initial value of the outcome being evaluated. We hypothesise that, in such settings, established methods for detecting IDEs will struggle to discriminate these from regression to the mean.

Methods: Using simulated datasets of weight-loss intervention programmes that recruit according to initial body mass index, we explore the reliability of Oldham's method and multilevel modelling (MLM) to detect IDEs.

Results: In datasets simulated with no IDE, Oldham's method and MLM yield Type I error rates $> 90\%$, confirming that threshold selection/truncation leads to bias due to regression to the mean. Type I error rates return close to 5% for both methods when a control group is introduced.

Conclusions: Oldham's method and MLM can robustly detect IDEs in this setting, but only if analyses incorporate a control group for comparison.

ARTICLE HISTORY
Received 29 March 2019
Accepted 28 December 2019

KEYWORDS
Intervention differential effect; mathematical coupling

Background

In longitudinal studies, researchers may be interested in detecting the presence of intervention differential effects (IDE). Intervention differential effects occur when the effect of an intervention on an outcome depends on the initial value of that outcome. Exploring IDEs in the analysis of change is akin to exploring the relationship between 'baseline' and

CONTACT Lucy Beggs lucy.beggs@nice.org.uk National Institute for Health and Care Excellence, Manchester, M1 4BT, United Kingdom

'change'; and statistical analysis of this relationship has been shown to face several challenges [1]. IDEs manifest as a change in the variance of outcome measures over time, with either a 'fanning out' or 'fanning in' of values (Figure 1) [1]. Recognising IDEs may have substantial clinical importance; for example, detecting the presence of IDEs may identify groups of patients most likely to benefit from an intervention. This article focuses on the challenges with established methods for detecting IDEs in samples that are selected based on 'high' initial values of the outcome variable. Using simulated datasets, we demonstrate that established methods do not perform robustly in these samples due to regression to the mean. We then propose an adaption to the methods and demonstrate that this restores their ability to identify IDEs.

Throughout the article, weight-loss studies are used as a motivating example because (i) weight-loss programs generally recruit based on high initial values of body mass index (BMI); and (ii) participants with a high BMI often lose more weight following an intervention than those with a lower initial BMI (an example of an IDE). Although this article focuses on weight-loss studies, our findings have relevance to all studies that explore the presence of an IDE in samples selected on the basis of the initial value of the outcome variable.

Current methods for detecting IDE

A search of the literature identified three established statistical approaches commonly used to detect IDEs: '*adjusting for baseline*'; *Oldham's method*; and *multilevel modelling*.

Adjusting for baseline

'Adjusting for baseline' to explore the presence of an IDE involves correlating or regressing change with baseline. In this seemingly intuitive approach, a significant correlation or baseline regression coefficient is inferred to indicate both the presence and strength of an IDE. However, Oldham [2] highlighted that change is *derived* from baseline; meaning that the correlation or regression model faces mathematical coupling. Mathematical coupling

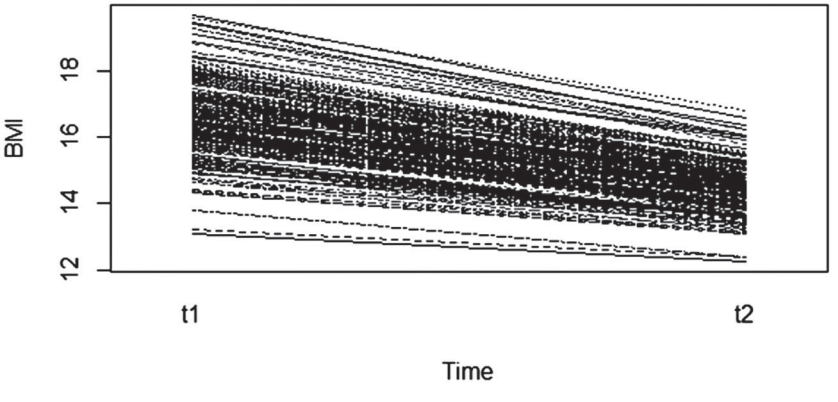


Figure 1. Simulated example of an IDE in which participants with high initial BMI have a greater intervention effect than participants with low initial BMI, leading to a reduction in standard deviation between measurements, manifesting as a 'fanning-in' of measurements over time.

occurs where '*one variable directly or indirectly contains the whole or part of another, and the two variables are then analysed using correlation or regression*' (1). It distorts the null hypothesis being tested, which may lead to incorrect estimates of the relationship between change and baseline. Tu and Gilthorpe [1] comprehensively outlined the problems of 'adjusting for baseline' in analyses of change. However, the erroneous practice of regressing change on baseline continues to be advocated as a robust method for detecting IDEs [3–6].

Oldham's method

Instead of identifying IDEs by adjusting for baseline, Oldham [2] proposed a method in which change is correlated with the average values of baseline and follow-up measurements; equivalent to a simple regression of change on average, and an approach that Tu and Gilthorpe [1] confirmed nullifies the adverse effects of mathematical coupling. Oldham's method works by identifying any change in the variance of the outcome measure from baseline to follow-up, detecting the 'fanning out' or 'fanning in' of measurements that emerge from an IDE. As such, Oldham's method is a well-established tool for detecting IDEs [7–9].

Multilevel modeling

Blance et al. [10] proposed that a post-intervention IDE can be estimated using a multilevel model, with time centered around zero as the principal covariate. Any non-zero covariance between the random intercept and the random slope implies the presence of an IDE. When only two longitudinal measures are involved (i.e. the pre-/post-intervention measurements), the estimated correlation is the same as Oldham's correlation. However, an advantage of multilevel modeling is that it allows for multiple repeated measures, and for the inclusion of additional covariates.

Hypothesis: regression to the mean (RTM) affects methods' ability to identify IDEs

As initially described by Galton [11], it is now well documented that longitudinal studies which select on the basis of high initial values will encounter regression to the mean (RTM). RTM implies that, following an extreme random event, the next random event is likely to be less extreme. In a sample selected on high initial value of the outcome variable, RTM leads to a 'fanning out' of values over time. This is because subsequent measures are likely to be less extreme and thus closer to the mean of the population from which they were selected (Figure 2). Crucially, this 'fanning out' is unrelated to any intervention and is entirely a consequence of selecting a sample based on a high initial value of the outcome variable. This is illustrated in the Appendix using approximated theoretical properties for a sample of repeated measures of BMI obtained by truncating a population of repeated BMI values *in the absence of an intervention* (i.e. assuming no IDE).

We thus hypothesise that, in samples selected based on an initial threshold (i.e. the initial value of the outcome variable), because IDEs and RTM both manifest as a change in the variance of measurements over time, identifying the presence of an IDE in samples selected on the basis of a high initial value of the outcome becomes challenging; and

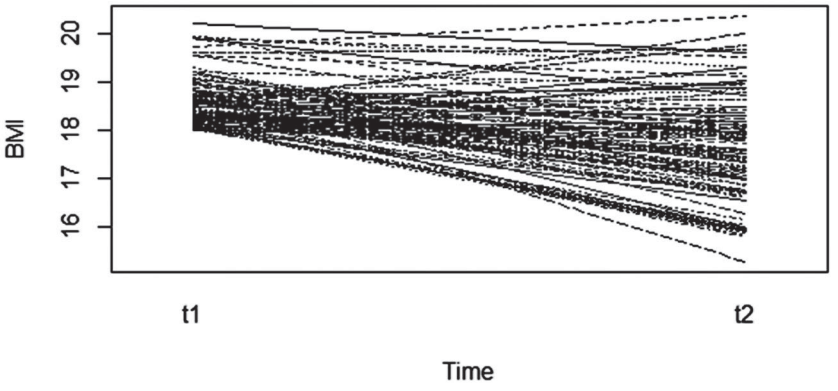


Figure 2. Simulated example of RTM, in which a sample is selected on the basis of high initial BMI. Over time, and independent of any intervention, the standard deviation of measurements increases due to RTM. This manifests as a 'fanning-out' of measurements over time.

that both Oldham's method and multilevel modeling will be unable to discriminate a genuine IDE from RTM. We use simulation methods to demonstrate this. We also reiterate the problem of mathematical coupling when an IDE is investigated by regressing change on baseline. Finally, we demonstrate that Oldham's method and multilevel modeling can robustly identify an IDE in the presence of RTM, but only if a control group is available.

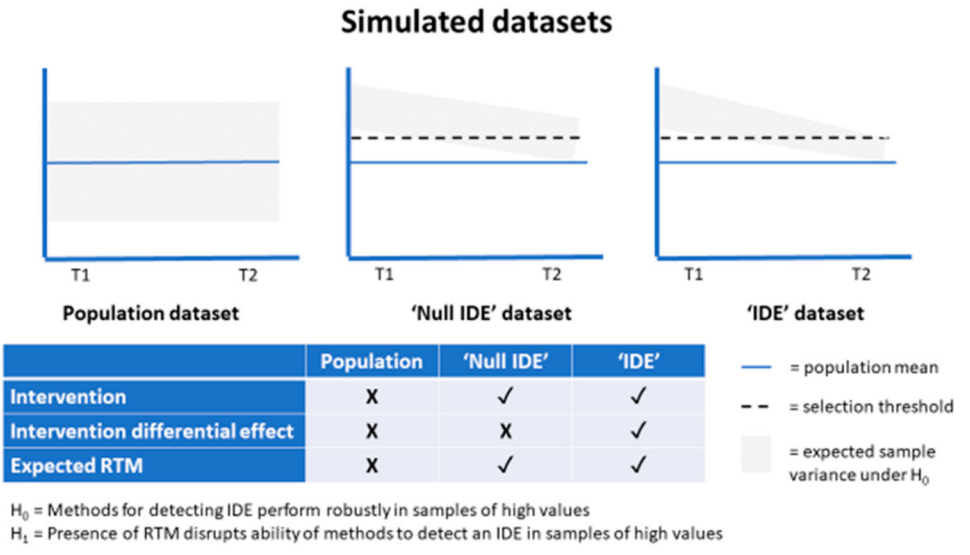
Methods

Simulation

BMI data for 5,000 male adults aged 25–34 were simulated, informed by data on obesity generated by the Health Survey for England [12]. Datasets were simulated under three hypothetical scenarios:

(i) a 'no intervention' control group with a mean BMI of 26.4 kg/m^2, and standard deviation (sd) of 5 kg/m^2 at baseline and at follow-up;
(ii) a 'null IDE' weight-loss program intervention group with a mean reduction in BMI of 4.4 kg/m^2, and no change in sd from baseline to follow-up; and,
(iii) a 'true IDE' weight-loss program intervention group with a mean reduction in BMI of 4.4 kg/m^2, and where participants with a high initial BMI lost more weight during the intervention period than participants with a lower initial BMI, such that BMI 'fanned-in' by 1.0 kg/m^2 to yield a sd of 4.0 kg/m^2.

These datasets were simulated to represent the population distribution of BMIs for adult males aged 25–34yrs. To emulate the sample recruitment process of a weight-loss program, a 'study weight-selection criterion' of BMI $\geq$30kg/m^2 (i.e. at or above the contemporary definition of clinical obesity) was then applied to each of the three datasets, and a random sample of 500 individuals was selected for each dataset. This produced 'control' (i), 'null IDE' (ii) and 'true IDE' samples (iii) selected on the basis of a high initial value of the outcome variable BMI (see Figure 3) – each of which can be thought of as 'truncated' samples



	Population	'Null IDE'	'IDE'
Intervention	X	✓	✓
Intervention differential effect	X	X	✓
Expected RTM	X	✓	✓

Figure 3. Schematic of the simulated truncated datatsets for each of the 3 scenarios: (i) 'no intervention' control; (ii) 'null IDE'; and (iii) 'true IDE'.

Table 1. Average properties of the simulated population and truncated sample datasets.

	Population	Study Sample
	kg/m^2	
Baseline mean BMI	26.4	32.9
Post intervention mean BMI – null IDE	22.0	27.6
Post intervention mean BMI – true IDE	22.0	26.4
Baseline BMI standard deviation	5.0	2.4
Post intervention BMI sd – null IDE	5.0	3.3
Post intervention BMI sd – true IDE	4.0	2.7
	Correlation	
Baseline / follow-up BMI association	0.85	0.62

of the population distribution of BMI values. This meant we had simulated population and truncated sample datasets for each of the 3 scenarios, resulting in a total of 6 datasets. By selecting samples based on a high initial value of the outcome variable BMI, the effects of RTM (which creates a 'fanning-out' of values) were introduced to all 3 truncated sample datasets. Additionally, in the truncated true IDE scenario (iii), the impact of the IDE (simulated as a 'fanning-in') was present *alongside* RTM. The simulated average parameter values for each of the population and truncated samples are summarised in Table 1. Simulations are detailed in the Appendix, together with sensitivity analyses that explored the uncertainty surrounding the choice of parameters. All simulations were undertaken in R, version 3.5.0 [13].

Statistical analyses

The 3 established methods used to detect an IDE were tested in both the population and the truncated sample datasets. To *adjust for baseline*, change in BMI was correlated with

Table 2. Median values and 95% credible intervals (95% CrI: 2.5% to 97.5% quantiles) of IDE correlations, and the proportion of associated p-values ≤ 0.05 (i.e. the Type I error rate under the null of no IDE; or the statistical power for a true IDE) for the analyses of population and truncated sample datasets, with 'adjusting for baseline', Oldham's method and MLM applied separately to each of the 'no intervention' control, 'null IDE', and 'true IDE' scenarios.

	Population		Study Subsample	
Method / Dataset	IDE Correlation (95% CrI)	% Statistically Significant	IDE Correlation (95% CrI)	% Statistically Significant
Adjusting for baseline				
Control	−0.27 (−0.30, −0.25)	100.0	−0.14 (−0.22, −0.05)	86.3
Null IDE	−0.27 (−0.30, −0.25)	100.0	−0.14 (−0.22, −0.05)	86.2
True IDE	−0.60 (−0.62, −0.59)	100.0	−0.35 (−0.42, −0.27)	100.0
Oldham's Method				
Control	0.00 (−0.03, 0.03)	5.3	0.38 (0.30, 0.46)	100.0
Null IDE	0.00 (−0.03, 0.03)	4.8	0.38 (0.30, 0.46)	100.0
True IDE	−0.39 (−0.42, −0.37)	100.0	0.13 (0.03, 0.22)	77.2
MLM (separate groups)				
Control	0.00 (−0.03, 0.03)	5.3	0.38 (0.30, 0.46)	100.0
Null IDE	0.00 (−0.03, 0.03)	4.8	0.38 (0.30, 0.46)	100.0
True IDE	−0.39 (−0.42, −0.37)	100.0	0.13 (0.03, 0.22)	77.2

baseline BMI to yield a coefficient (and associated p-value) equivalent to the standardised regression of change on baseline. *Oldham's method* was used to derive the correlation (and associated p-value) between change and the mean of baseline and follow-up measurements. The *multilevel model* assigned baseline and follow-up BMI measurements to level-1, and individuals to level-2; and centered the covariate measurement occasion (i.e. time) on its mean, while estimating the covariance between random intercept and random slope to derive a correlation (and associated p-value) – equivalent to Oldham's method [10]. All estimates of correlations and p-values were derived using restricted maximum likelihood (REML), and were stored for each iteration.

Empirical distributions of correlation coefficients (or correlation differences) were summarised using a median point estimate and 95% credible interval (95% CrI: 2.5% to 97.5% quantiles). Associated empirical distributions of each set of p-values were used to derive the proportions of p-values that were $\leq$0.05. These proportions indicate Type I error rates for the control (i) and null IDE scenarios (ii), while revealing the statistical power to detect the simulated 'true' IDE. A robust method is expected to find around 5% of iterations to be significant in the control and null IDE scenarios; a higher proportion of Type I error rate indicating that the method under question is not robust. All analyses were undertaken in R, version 3.5.0 [13], with multilevel models developed in MLwiN, version 3.01 [14] from within R using the R2MLwiN package [15].

Results

Table 2 summarises the results of all analyses undertaken on both the simulated populations and the truncated samples. These analyses estimated the IDE and associated p-values separately for the control (i), 'null IDE' (ii) and 'true IDE' scenarios (iii).

Population datasets

As expected, Oldham's method and MLM yielded identical results when applied to the population control and null IDE scenarios. This is because both methods detect IDEs by assessing the change in standard deviation from baseline to follow-up, which in this instance was zero in both scenarios because the control scenario (i) displayed no overall mean change and no change in sd, while the null IDE intervention scenario (ii) displayed a mean change in BMI but no change in sd. Adjusting for baseline also yielded identical estimates between control and null IDE scenarios. However, these estimates were consistently biased, with high Type I error rates, for the control and null IDE scenarios arising because adjusting for baseline fails to account for mathematical coupling; and since this is shown to be highly biased, it is not meaningful to interpret the IDE scenario or truncated sample evaluations as indicating the presence of a genuine IDE.

Truncated sample datasets

Oldham's method and MLM also yielded equivalent results for the control group (i) and null IDE intervention group (ii) within the truncated sample datasets. However, in both groups the estimated IDE was severely biased by RTM in the truncated sample dataset. In the null IDE group, Type I error rates were 100% for both Oldham's method and MLM. This clearly demonstrates how the recruitment of a study sample according to high initial value of the outcome (as with weight-management studies) can bias the estimated IDE for both these methods. In the scenario simulated to have a 'true' IDE of −0.39, estimation within the sampled dataset was 0.13. In this instance, the presence of RTM biased the methods to a sufficient extent to reverse the predicted direction of effect for the IDE; instead of a 'fanning-in' IDE, both Oldham's method and MLM would estimate a 'fanning-out' IDE. This is due to the dominant adverse effects of RTM which overwhelm the 'true' IDE (the impact of RTM working in the opposite direction to the simulated IDE).

An alternative approach

Our statistical analysis demonstrates that 'adjusting for baseline' is biased even in the population samples (i.e. in the absence of RTM), and so should not be used to identify IDEs. We have also shown Oldham's method and MLM cannot robustly identify IDEs in truncated samples (i.e. wherever RTM is present). We now propose an alternative approach for identifying IDEs in samples selected on the basis of high initial values. This approach is predicated upon the simple fact that IDEs occur only after an intervention, and so can only genuinely occur in an intervention group; whilst RTM would be present in both intervention and control groups of any truncated sample. We exploited this premise to develop an approach that isolates IDEs from RTM by contrasting analyses in an intervention group using data from a control group.

To attempt to isolate an IDE from RTM using both Oldham's method and MLM, the datasets from the 'true IDE' (iii) and 'null IDE' (ii) intervention scenarios were contrasted with the dataset from the control scenario (i). To contrast the scenario datasets using Oldham's method, pairwise contrasts for each method were obtained using Fisher's z-transformation [16] and student's t-test to yield associated p-values for both population

and threshold selected/truncated samples. To contrast the scenario datasets using a multilevel model, control and intervention scenario data were combined and analysed using two nested multilevel models: one where the covariance of random intercept and random slope for each scenario was independently estimated (i.e. allowed to differ, as would be appropriate were an IDE to exist), and one where this covariance was forced to be identical for both scenarios (as would be appropriate were no IDE to exist). The correlations derived in the first multilevel model were differenced, and the nested models were evaluated using the likelihood ratio test (and restricted maximum likelihood method) [17] to yield a *p*-value. All correlation differences and *p*-values were stored.

Table 3 summarises the IDEs estimated by differencing correlations obtained when contrasting control and intervention scenario datasets using Fisher's z-transformation and t-test, as well as the alternative approach of using the likelihood ratio test [17] to contrast nested multilevel models. Oldham's method and MLM overcame the adverse impacts of RTM when a control group was introduced to evaluate evidence for an IDE through statistical testing – Type 1 error rates for all methods being in the region of 5%. Both methods correctly estimated IDE correlations for the population dataset. However, no method could correctly estimate the magnitude of the IDE correlations in these samples, even though Oldham's method and MLM were less biased than adjusting for baseline within the truncated sample dataset.

Fisher's z-transformation and t-test appeared to be slightly less robust when evaluating separate group applications of Oldham's method or MLM (6.7% Type 1 error) compared to the likelihood ratio test for nested models of combined group MLMs (5.2% Type 1 error). However, when this was explored with sensitivity analyses (Appendix), there was no consistency in the small differences observed between each approach. In contrast, for

Table 3. The proportion of associated *p*-values ≤ 0.05 (i.e. the Type I error rate under the null of no IDE; or the statistical power for a true IDE) for the analyses of population and truncated sample datasets, with 'adjusting for baseline', Oldham's method and two MLM approaches: (i) using separate models to contrast correlations between control and intervention groups using Fisher's z-transformation and t-test; and (ii) using combined control and intervention group models with the likelihood ratio test of nested models for random intercept / slope covariances allowed to differ, or constrained to be equal, across groups.

	Population		Study Subsample	
Method / Dataset	Difference in IDE Correlations (95% CrI)	% Statistically Significant	Difference in IDE Correlations (95% CrI)	% Statistically Significant
Adjusting for baseline				
Control vs. No IDE	0.00 (−0.04, 0.04)	4.6	0.00 (−0.12, 0.12)	5.0
Control vs. True IDE	−0.33 (−0.36, −0.30)	100.0	−0.21 (−0.33, −0.09)	93.9
Oldham's Method				
Control vs. Null IDE	0.00 (−0.04, 0.04)	4.7	0.00 (−0.11, 0.11)	6.7
Control vs. True IDE	−0.39 (−0.43, −0.36)	100.0	−0.26 (−0.38, −0.13)	98.8
MLM (separate groups)				
Control vs. Null IDE	0.00 (−0.04, 0.04)	4.7	0.00 (−0.11, 0.11)	6.7
Control vs. True IDE	−0.39 (−0.43, −0.36)	100.0	−0.26 (−0.38, −0.13)	98.8
MLM (combined groups)				
Control vs. Null IDE	0.00 (−0.04, 0.04)	4.8	0.00 (−0.11, 0.11)	5.2
Control vs. True IDE	−0.39 (−0.43, −0.36)	100.0	−0.26 (−0.38, −0.13)	99.8

the simulated IDE the statistical power attained was highest for the likelihood ratio test (99.8%), and slightly less for contrasting Oldham's method or separate MLMs (98.8%). This pattern of relative strengths in statistical power was found to be consistent (Appendix).

Discussion

'Adjusting for baseline' consistently identified large and statistically significant IDEs in datasets deliberately simulated to have no IDE. This was even the case in the absence of RTM, illustrating how this approach is *always* inappropriate, as previously stated [1,2].

In the 'null IDE' scenario, Oldham's method and multilevel modeling yielded reasonable Type 1 error rates for the population dataset, correctly indicating that there is no IDE. However, both methods gave rise to very high Type 1 error rates of 100% for the truncated sample datasets, implying the presence of an IDE. As the 'null IDE' datasets were simulated not to have an IDE, our analysis demonstrates how both methods mistake the effects of RTM as an IDE.

By introducing a control group that is truncated in the same way as the study sample, and is therefore prone to the same impacts of RTM as the intervention group, we restored unbiased Type I error rates for the 'null IDE' scenario. Contrasting correlations in the control and intervention groups using Oldham's method and MLM gave the expected Type I error rate for all methods. This indicates that our alternative approach is robust, and provides a solution to the challenges of detecting IDEs in samples selected on the basis of a high initial value of the outcome variable. The MLM approach of analysing combined control and intervention data, then testing nested models with and without constrained covariances between groups using the likelihood ratio test, was the most powerful overall of all methods considered. The key point, however, was that for either method to be effective, a control group was required.

When seeking to estimate the *magnitude* of the IDE estimated by Oldham's method and/or MLM, considering the separately derived correlations for control and intervention groups was robust within the population dataset.

Since our datasets were simulated with the 'true' IDE operating in the *opposite* direction to the RTM, the estimated IDE correlation for the IDE scenario was sufficiently biased to have its sign reversed (Table 2: 'true' IDE correlation was −0.39 but estimated as 0.13 in the truncated sample datasets using either Oldham's method or MLM). If the 'true' IDE were to operate in the *same* direction as the effects of RTM we hypothesise that the overall estimated IDE correlation would not be reversed, but instead enhanced. For either direction of an IDE, the impacts of RTM are likely to be substantial and will prevent reliable estimation of the magnitude of a genuine IDE. Thus, even with a control group, it is not possible to estimate the IDE correlation within a truncated sample dataset; it is only possible to 'test' if there is sufficient evidence that an IDE is present. The *direction* of IDE must then be ascertained by inspecting intervention and control groups to establish in which the 'fanning in' or 'fanning out' is greater.

Focusing on testing for the presence of an IDE, rather than estimating its magnitude, places the emphasis on study size and statistical power. For an adult weight management study with a population IDE of 1 kg/m^2 sd reduction (corresponding to 0.6 kg/m^2 sd reduction in the truncated sample dataset), Oldham's method and MLM made use of the control group to provide good statistical power (even though Oldham's method utilises only the

first and last measure of a longitudinal dataset). Sensitivity analysis shows that smaller intervention effects (e.g. a mean BMI reduction of just 2 kg/m^2), smaller IDEs (e.g. a sd reduction of just 0.2 kg/m^2) and stronger autocorrelation (e.g. rho = 0.95) all diminish the statistical power – power diminishing to around 36% for the Fisher's z-transformation and to around 39% for the likelihood ratio test in the most stringent scenario (see Appendix Table A4b). To improve statistical power, MLM can make good use of any additional longitudinal measures available. MLM can also be used to explore longitudinal data with *multiple* pre-intervention measures provided there is at least one post-intervention measure. The approach involves estimating separate IDEs for the pre-intervention period and the post-intervention period, such that the study sample operates as its own control.

Greater serial autocorrelation in the outcome leads to weaker adverse impacts of RTM. Furthermore, the extent of RTM in the study dataset depends on the selection threshold/level of truncation adopted. The extent of bias in any IDE estimate thus depends on the frequency of the longitudinal measures adopted and on the selection criteria used to truncate the population sample. Both serial autocorrelation and sample selection are study- and context-specific; their impact on different studies were explored through sensitivity analyses (see Appendix). Unsurprisingly, RTM had greater adverse impacts on the validity of Oldham's method and MLM in samples where the outcome exhibited lower serial autocorrelation, or where samples were derived from more extreme sample selection/truncation. Nevertheless, relatively high serial autocorrelation values (e.g. 0.9) and modest sampling thresholds/truncation (e.g. above/below the mean) yielded sizeable bias due to RTM, with Type 1 error rates $> 90\%$ for both Oldham's method and MLM under the null of no IDE.

A key assumption throughout this study is that the outcome in the control group is homoscedastic. This is not true in some contexts: for instance, weight-management programs in *children* will observe that weight is heteroscedastic in the absence of intervention due to children's underlying growth trajectories. While beyond the scope of this study, evaluating the robustness of models evaluating IDEs for outcomes known to exhibit homoscedasticity in the non-intervention population is an important area for further consideration.

The methodological findings of this paper are highly relevant to all studies exploring IDEs in samples recruited/truncated above (or below) a threshold. However, in the context of weight-loss studies, the implications are particularly important because studies recruiting participants with very high BMIs often do not recruit control participants. For example, in Janicke et al.'s [18] systematic review of randomised controlled trials, they excluded 54 of 278 papers selected for full-text review because they lacked appropriate control groups. Likewise, Benestad et al.'s [19] study noted a similar lack of adequate control groups in observational (i.e. non-experimental) weight-loss studies. For studies in samples recruited on high initial value that lack a control group, any subsequent claim of a relationship between change and baseline is at best biased, but is more likely to be meaningless or completely misleading.

Conclusion

This paper evaluates three established methods (adjusting for baseline, Oldham's method, and MLM) commonly used to evaluate IDEs in longitudinal analyses of change in samples selected on the basis of a high initial value of the outcome variable. We have shown that

'adjusting for baseline' performs poorly in multiple scenarios. We have demonstrated that Oldham's method and multilevel modeling perform robustly in samples that reflect the full population distribution of values, but these methods are not robust in samples that select based on high/low initial values of the outcome, due to the adverse impacts of regression to the mean. Both Oldham's method and multilevel modeling can nonetheless robustly detect the presence of IDEs in studies that select according to high initial value, provided contrasts are made between the intervention group and a control group. We have shown that it is not possible to estimate the *size* of an IDE; it is only possible to 'test' for the *presence* of an IDE. We have also concluded that the direction of any IDE must be ascertained by inspecting the 'fanning-in' or 'fanning-out' of outcome values within the intervention and control groups.

To conclude, a control dataset is needed to robustly identify IDEs in the presence of RTM. Failure to recognise the consequences of mathematical coupling and RTM are likely to lead to incorrect estimates of the relationship between change and baseline in samples based on truncated (high/low) initial values of the outcome variable.

Acknowledgement

All authors made substantial contributions to the conception and design of this study, and have been involved in drafting or revising the manuscript. Lucy Beggs, Rebecca Briscoe and Mark Gilthorpe have made substantial contributions to the analysis and interpretation of the data.

Disclosure statement

No potential conflict of interest was reported by the authors.

Funding

Mark Gilthorpe supported by The Alan Turing Institute; Fellowship Grant Number EP/N510129/1.

Notes on contributors

Lucy Beggs is Health Technology Assessment Advisor at the National Institute for Health and Care Excellence.

Rebecca Briscoe is a Public Health Registrar in Leeds.

Claire Griffiths is Reader in the Carnegie School of Sport at Leeds Beckett University.

George T. H. Ellison is Associate Professor of Epidemiology and Deputy Director of Leeds Institute for Data Analytics.

Mark S. Gilthorpe is Professor of Statistical Epidemiology and Fellow of The Alan Turing Institute.

Availability of data and materials

All data generated or analysed during this study can be replicated using the R code that is available at: https://github.com/VeetVoojagig/IDE.

ORCID

Lucy Beggs http://orcid.org/0000-0002-0574-6325

Rebecca Briscoe http://orcid.org/0000-0002-6611-9317
Claire Griffiths http://orcid.org/0000-0002-2588-1022
George T. H. Ellison http://orcid.org/0000-0001-8914-6812
Mark S. Gilthorpe http://orcid.org/0000-0001-8783-7695

References

[1] Tu Y-K, Gilthorpe MS. Revisiting the relation between change and initial value: a review and evaluation. Stat Med. 2007;26(2):443–457.
[2] Oldham PD. A note on the analysis of repeated measurements of the same subjects. J Chronic Dis. 1962;15:969–977.
[3] Guessous I, McClellan W, Kleinbaum D, et al. Serum 25-hydroxyvitamin D level and kidney function decline in a Swiss general adult population. Clin J Am Soc Nephr. 2015;10(7):1162–1169.
[4] Squara P. Mathematic coupling of data: a frequently misused concept. Intens Care Med. 2008;34(10):1916–1921.
[5] Terluin B. Mathematical coupling does not account for the association between baseline severity and minimally important change values. J Clin Epidemiol. 2012;65(4):355–357.
[6] Takase B, Akima T, Uehata A, et al. Endothelial function as a possible significant determinant of cardiac function during exercise in patients with structural heart disease. Cardiol Res. 2009;2009:1–9.
[7] Snider S, Quisenberry A, Bickel W. Order in the absence of an effect: identifying rate-dependent relationships. Behav Proc. 2016;127:18–24.
[8] Mishra E, Corcoran J, Hallifax R, et al. Defining the minimal important difference for the visual analogue scale assessing dyspnea in patients with malignant pleural effusions. PLoS One. 2015;10(4):e0123798.
[9] Baldinger P, Kranz G, Haeusler D, et al. Regional differences in SERT occupancy after acute and prolonged SSRI intake investigated by brain PET. Neuroimage. 2014;88:252–262.
[10] Blance A, Tu Y-K, Gilthorpe MS. A multilevel modelling solution to mathematical coupling. Stat Methods Med Res. 2005;14(6):553–565.
[11] Galton F. Regression toward mediocrity in hereditary stature. J Anthropol Inst Grt Br Ir. 1886;15:246–263.
[12] NHS Digital. *Health Survey for England*, 2016. [2017; cited 2019 Oct]. Available from: https://digital.nhs.uk/catalogue/PUB30169.
[13] R Core Team. R: A language and environment for statistical computing. Vienna: R Foundation for Statistical Computing; 2018; Available from: https://www.r-project.org/.
[14] Charlton C, Rasbash J, Browne WJ, et al. MLwin version 3.00. Centre for multilevel modelling. Bristol: University of Bristol; 2017.
[15] Zhang Z, Parker RM, Charlton CM, et al. R2MLwiN: A package to run MLwiN from within R. J Stat Softw. 2016;72(10):1–43.
[16] Fisher RA. Statistical methods for research Workers. New York: Hafner; 1958.
[17] Neyman J, Pearson ES. On the problem of the most efficient tests of statistical hypotheses. Philos Trans R Soc A Math Phys Eng Sci. 1933;231(694–706):289–337.
[18] Janicke DM, Steele RG, Gayes LA, et al. Systematic review and meta-analysis of comprehensive behavioral family lifestyle interventions addressing pediatric obesity. J Pediatr Psychol. 2014;39(8):809–825.
[19] Benestad B, Lekhal S, Småstuen MC, et al. Camp-based family treatment of childhood obesity: randomised controlled trial. Arch Dis Child. 2017;102(4):303–310.
[20] Aitkin M. Correlation in a singly truncated bivariate normal distribution. Psychometrika. 1964;29:263–270.
[21] Mills JP. Table of the ratio: area to bounding ordinate, for any portion of normal curve. Biometrika. 1926;18:395–400.

Appendix

Theoretical illustration of regression-to-the-mean

For a population of adults with a baseline measure of body mass index (BMI) denoted by x, and follow-up BMI a short interval later (e.g. a few weeks) denoted by y, we assume x and y follow a bivariate normal distribution:

$$\text{BMI} = N(\text{M}, \Sigma), \text{where M} = \begin{bmatrix} \mu_x \\ \mu_y \end{bmatrix} \text{and } \Sigma = \begin{bmatrix} \sigma_x^2 & \rho_{xy} \\ \rho_{xy} & \sigma_y^2 \end{bmatrix},$$

for baseline (μ_x) and follow-up (μ_y) mean BMI; baseline (σ_x) and follow-up (σ_y) BMI standard deviations; and Pearson correlation coefficient (ρ_{xy}) between baseline and follow-up BMI.

If we sample baseline BMI above a threshold, t_x, for corresponding z-score, $z = (t_x - \mu_x)/\sigma_x$, we obtain a lower-bounded, singularly truncated bivariate distribution of BMI, denoted $\{\check{x}, \check{y}\}$, with properties [20]:

$$\sigma_{\check{x}} = \sqrt{1 + \frac{z}{R(z)} - \frac{1}{R^2(z)}},$$

$$\sigma_{\check{y}} = \sqrt{1 + \frac{\rho_{xy}^2 z}{R(z)} - \frac{\rho_{xy}^2}{R^2(z)}},$$

$$\rho_{\check{x}\check{y}} = \rho_{xy}\sqrt{\frac{R^2(z) + zR(z) - 1}{R^2(z) + \rho_{xy}^2(zR(z) - 1)}},$$

where

$$R(z) = e^{\frac{x^2}{2}} \int_x^\infty e^{\frac{-t^2}{2}} dt$$

is known as the Mills' ratio [21].

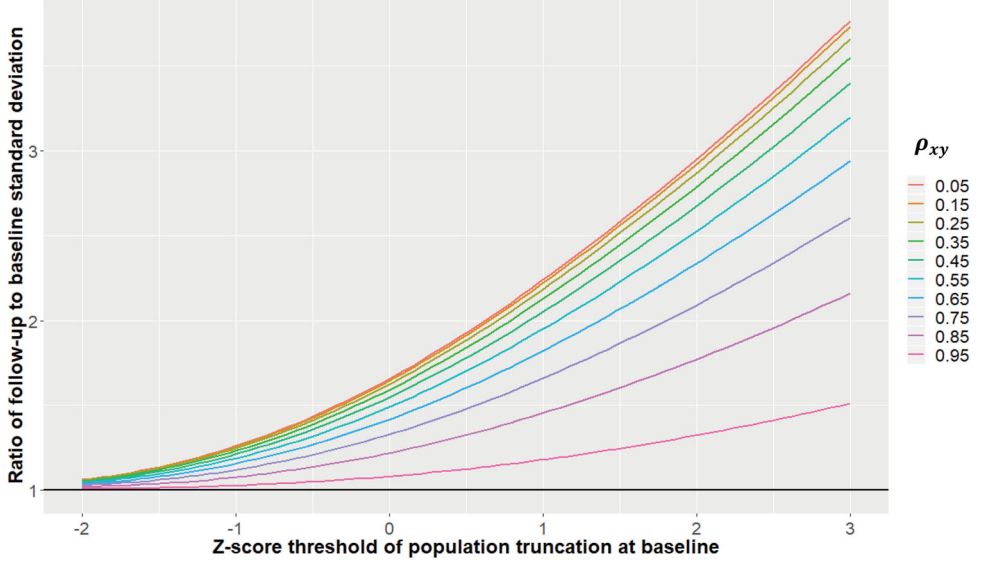


Figure A1. The estimated ratio of follow-up standard deviation ($\sigma_{\check{y}}$) to baseline standard deviation ($\sigma_{\check{x}}$) in a lower-bounded, singularly truncated bivariate distribution of adult body mass index measures (BMI) sampled for baseline BMI above threshold z-scores (z), for a population whose bivariate correlation of baseline with follow-up BMI measures (ρ_{xy}) range between 0.05 and 0.95.

Under the null assumption of no intervention differential effect (IDE), i.e. there no exogenous influences that affect the mean or standard deviation of BMI measures, then $\sigma_x = \sigma_y$ and the ratio $\sigma_y/\sigma_x \equiv 1$ for the population. In the lower-bounded, singularly truncated bivariate distribution, however, the corresponding ratio, $\sigma_{\check{y}}/\sigma_{\check{x}}$, is no longer unity but varies according to ρ_{xy} and z, as demonstrated in Figure A1.

For any truncation, $\sigma_{\check{y}}/\sigma_{\check{x}} > 1$, and the extent of biased assessment of an IDE increases as the truncated sample dataset becomes more restricted to higher baseline thresholds. Bias is more extreme for smaller bivariate correlations between baseline and follow-up BMI measures. Even for a correlation of 0.95, truncation above the population mean (z-score $= 0$) yields a $\sigma_{\check{y}}/\sigma_{\check{x}} = 1.082$, indicating that follow-up BMI standard deviation is 8.2% larger than baseline BMI standard deviation (i.e. a fanning out of BMI from baseline), which is entirely due to regression to the mean. For sample truncation above BMI ≥ 30, i.e. selecting those formally identified as 'obese' (z-score $= 0.72$), follow-up BMI will have a standard deviation 14.8% larger than baseline BMI standard deviation due entirely to regression to the mean.

Sensitivity analyses

To undertake sensitivity analyses, three additional scenarios were considered to complement the default (Sim 1) in the main text (see Table A1). To expedite matters, simulations were repeated 1,000 times only (this took a full day of processing time for each different scenario) as performing 10,000 simulations, albeit more precise in estimating Type 1 errors and statistical power, would not affect any conclusions drawn. Corresponding findings for each simulation are summarised in Tables A2a to A4b.

Table A1. The simulation parameters considered for each scenario.

	Sim 1	Sim 2	Sim 3	Sim 4
		kg/m^2		
Population baseline mean BMI	26.4	26.4	26.4	26.4
Population follow-up mean BMI	22.0	23.0	23.0	24.4
Population baseline standard deviation	5.0	4.0	4.0	4.0
Population follow-up true IDE standard deviation	4.0	3.5	3.5	3.8
Recruitment BMI threshold	30.0	30.0	26.4	26.4
		Correlation		
Population baseline-follow-up BMI correlation	0.85	0.95	0.90	0.95

Table A2a. Median and 95% credible interval (95% CrI: 2.5% to 97.5% quantiles) of IDE correlations and the proportion of associated *p*-values $\leq$ 0.05 (Type I error rate under the null of no IDE or statistical power for a true IDE) for the analyses of BMI population and study subsample in simulation 2 (Table A1), with adjustment for baseline, Oldham's method and multilevel modeling applied separately to control group, intervention group with no IDE, and intervention group with a true IDE.

	Population		Study Subsample	
Method / Dataset	IDE Correlation (95% CrI)	% Statistically Significant	IDE Correlation (95% CrI)	% Statistically Significant
Adjusting for baseline				
Control	−0.16 (−0.18, −0.13)	100.0	−0.07 (−0.16, 0.01)	38.0
Null IDE	−0.16 (−0.18, −0.13)	100.0	−0.07 (−0.16, 0.01)	38.2
True IDE	−0.53 (−0.54, −0.51)	100.0	−0.27 (−0.36, −0.19)	100.0
Oldham's Method				
Control	0.00 (−0.03, 0.03)	5.9	0.26 (0.18, 0.34)	100.0
Null IDE	0.00 (−0.03, 0.03)	4.2	0.26 (0.18, 0.34)	100.0
True IDE	−0.39 (−0.42, −0.37)	100.0	0.04 (−0.05, 0.13)	15.5
MLM (separate groups)				
Control	0.00 (−0.03, 0.03)	5.9	0.26 (0.18, 0.34)	100.0
Null IDE	0.00 (−0.03, 0.03)	4.2	0.26 (0.18, 0.34)	100.0
True IDE	−0.39 (−0.42, −0.37)	100.0	0.04 (−0.05, 0.13)	15.5

Table A2b. The proportion of associated *p*-values $\leq$0.05 (Type I error rate under the null of no IDE and statistical power for true IDE) for the analyses of BMI population and study subsample in simulation 2 (Table A1), with adjusting for baseline, Oldham's method and two multilevel modeling approaches: (i) separate models contrasting correlations between control and intervention groups using Fisher's z-transformation and t-test; and (ii) combined control and intervention group models with the likelihood ratio test of nested models for random intercept / slope covariances allowed to differ or constrained to be equal across groups.

	Population		Study Subsample	
Method / Dataset	Difference in IDE Correlations (95% CrI)	% Statistically Significant	Difference in IDE Correlations (95% CrI)	% Statistically Significant
Adjusting for baseline				
Control vs. Null IDE	0.00 (−0.04, 0.04)	4.3	0.00 (−0.12, 0.12)	4.1
Control vs. True IDE	−0.37 (−0.40, −0.34)	100.0	−0.20 (−0.31, −0.08)	90.6
Oldham's Method				
Control vs. Null IDE	0.00 (−0.04, 0.04)	4.9	0.00 (−0.12, 0.11)	4.9
Control vs. True IDE	−0.40 (−0.43, −0.36)	100.0	−0.22 (−0.34, −0.10)	94.5
MLM (separate groups)				
Control vs. Null IDE	0.00 (−0.04, 0.04)	4.9	0.00 (−0.12, 0.11)	4.9
Control vs. True IDE	−0.40 (−0.43, −0.36)	100.0	−0.22 (−0.34, −0.10)	94.5
MLM (combined groups)				
Control vs. Null IDE	0.00 (−0.04, 0.04)	4.9	0.00 (−0.12, 0.11)	5.2
Control vs. True IDE	−0.40 (−0.43, −0.36)	100.0	−0.22 (−0.34, −0.10)	97.5

Table A3a. Median and 95% credible interval (95% CrI: 2.5% to 97.5% quantiles) of IDE correlations and the proportion of associated *p*-values ≤ 0.05 (Type I error rate under the null of no IDE or statistical power for a true IDE) for the analyses of BMI population and study subsample in simulation 3 (Table A1), with adjusting for baseline, Oldham's method and multilevel modeling applied separately to control group, intervention group with no IDE, and intervention group with true IDE.

	Population		Study Subsample	
Method / Dataset	IDE Correlation (95% CrI)	% Statistically Significant	IDE Correlation (95% CrI)	% Statistically Significant
Adjusting for baseline				
Control	−0.22 (−0.25, −0.20)	100.0	−0.13 (−0.22, −0.05)	86.4
No IDE	−0.22 (−0.25, −0.20)	100.0	−0.13 (−0.23, −0.05)	86.0
True IDE	−0.49 (−0.51, −0.47)	100.0	−0.32 (−0.39, −0.24)	100.0
Oldham's Method				
Control	0.00 (−0.03, 0.03)	5.9	0.23 (0.14, 0.31)	100.0
Null IDE	0.00 (−0.03, 0.03)	4.2	0.23 (0.14, 0.31)	99.8
True IDE	−0.29 (−0.32, −0.27)	100.0	0.01 (−0.07, 0.11)	7.8
MLM (separate groups)				
Control	0.00 (−0.03, 0.03)	5.9	0.23 (0.14, 0.31)	100.0
Null IDE	0.00 (−0.03, 0.03)	4.2	0.23 (0.14, 0.31)	99.8
True IDE	−0.29 (−0.32, −0.27)	100.0	0.01 (−0.07, 0.11)	7.8

Table A3b. The proportion of associated *p*-values ≤ 0.05 (Type I error rate under the null of no IDE and statistical power for true IDE) for the analyses of BMI population and study subsample in simulation 3 (Table A1), with adjusting for baseline, Oldham's method and two multilevel modeling approaches: (i) separate models contrasting correlations between control and intervention groups using Fisher's z-transformation and t-test; and (ii) combined control and intervention group models with the likelihood ratio test of nested models for random intercept / slope covariances allowed to differ or constrained to be equal across groups.

	Population		Study Subsample	
Method / Dataset	Difference in IDE Correlations (95% CrI)	% Statistically Significant	Difference in IDE Correlations (95% CrI)	% Statistically Significant
Adjusting for baseline				
Control vs. Null IDE	0.00 (−0.04, 0.03)	4.5	0.00 (−0.12, 0.12)	4.6
Control vs. True IDE	−0.26 (−0.30, −0.23)	100.0	−0.18 (−0.30, −0.07)	87.3
Oldham's Method				
Control vs. Null IDE	0.00 (−0.04, 0.04)	4.9	0.00 (−0.12, 0.11)	4.6
Control vs. True IDE	−0.30 (−0.33, −0.26)	100.0	−0.21 (−0.33, −0.08)	91.3
MLM (separate groups)				
Control vs. Null IDE	0.00 (−0.04, 0.04)	4.9	0.00 (−0.12, 0.11)	4.6
Control vs. True IDE	−0.30 (−0.33, −0.26)	100.0	−0.21 (−0.33, −0.08)	91.3
MLM (combined groups)				
Control vs. Null IDE	0.00 (−0.04, 0.04)	4.9	0.00 (−0.12, 0.11)	4.3
Control vs. True IDE	−0.30 (−0.33, −0.26)	100.0	−0.21 (−0.33, −0.08)	94.9

Table A4a. Median and 95% credible interval (95% CrI: 2.5% to 97.5% quantiles) of IDE correlations and the proportion of associated *p*-values ≤ 0.05 (Type I error rate under the null of no IDE or statistical power for a true IDE) for the analyses of BMI population and study subsample in simulation 4 (Table A1), with adjusting for baseline, Oldham's method and multilevel modeling applied separately to control group, intervention group with no IDE, and intervention group with true IDE.

	Population		Study Subsample	
Method / Dataset	IDE Correlation (95% CrI)	% Statistically Significant	IDE Correlation (95% CrI)	% Statistically Significant
Adjusting for baseline				
Control	−0.16 (−0.18, −0.13)	100.0	−0.10 (−0.18, −0.01)	58.9
Null IDE	−0.16 (−0.18, −0.13)	100.0	−0.10 (−0.19, −0.01)	58.9
True IDE	−0.31 (−0.34, −0.29)	100.0	−0.19 (−0.27, −0.11)	99.6
Oldham's Method				
Control	0.00 (−0.03, 0.03)	5.9	0.16 (0.08, 0.25)	95.4
Null IDE	0.00 (−0.03, 0.03)	4.2	0.16 (0.07, 0.25)	94.0
True IDE	−0.16 (−0.19, −0.14)	100.0	0.06 (−0.03, 0.15)	26.7
MLM (separate groups)				
Control	0.00 (−0.03, 0.03)	5.9	0.16 (0.08, 0.25)	95.4
Null IDE	0.00 (−0.03, 0.03)	4.2	0.16 (0.07, 0.25)	94.0
True IDE	−0.16 (−0.19, −0.14)	100.0	0.06 (−0.03, 0.15)	26.7

Table A4b. The proportion of associated *p*-values ≤ 0.05 (Type I error rate under the null of no IDE and statistical power for true IDE) for the analyses of BMI population and study subsample in simulation 4 (Table A1), with adjusting for baseline, Oldham's method and two multilevel modeling approaches: (i) separate models contrasting correlations between control and intervention groups using Fisher's z-transformation and t-test; and (ii) combined control and intervention group models with the likelihood ratio test of nested models for random intercept / slope covariances allowed to differ or constrained to be equal across groups.

	Population		Study Subsample	
Method / Dataset	Difference in IDE Correlations (95% CrI)	% Statistically Significant	Difference in IDE Correlations (95% CrI)	% Statistically Significant
Adjusting for baseline				
Control vs. Null IDE	0.00 (−0.04, 0.04)	4.3	0.00 (−0.13, 0.12)	5.9
Control vs. True IDE	−0.16 (−0.20, −0.12)	100.0	−0.10 (−0.21, 0.03)	33.2
Oldham's Method				
Control vs. Null IDE	0.00 (−0.04, 0.04)	4.9	0.00 (−0.12, 0.12)	5.7
Control vs. True IDE	−0.16 (−0.20, −0.13)	100.0	−0.10 (−0.22, 0.03)	36.2
MLM (separate groups)				
Control vs. Null IDE	0.00 (−0.04, 0.04)	4.9	0.00 (−0.12, 0.12)	5.7
Control vs. True IDE	−0.16 (−0.20, −0.13)	100.0	−0.10 (−0.22, 0.03)	36.2
MLM (combined groups)				
Control vs. Null IDE	0.00 (−0.04, 0.04)	4.9	0.00 (−0.12, 0.12)	5.3
Control vs. True IDE	−0.16 (−0.20, −0.13)	100.0	−0.10 (−0.22, 0.03)	39.0