Re-expressing coefficients from regression models for inclusion in a meta-analysis

Matthew W. Linakis,[a,*] Cynthia Van Landingham,[b] Alessandro Gasparini[c], Matthew P. Longnecker[a]

[a] Ramboll U.S. Consulting, Raleigh NC 27612 U.S.A.

[b] Ramboll U.S. Consulting, Monroe LA 71201 U.S.A.

[c] Red Door Analytics AB, Stockholm, Sweden

*Corresponding Author:
Matthew W. Linakis
Ramboll, 3214 Charles B Root Wynd #130, Raleigh, NC 27612 U.S.A.
mlinakis@ramboll.com
Phone: (919) 987-3063

Abstract

Meta-analysis poses a challenge when original study results have been expressed in a non-uniform manner, such as when regression results from some original studies were based on a log-transformed key independent variable while in others no transformation was used. Methods of re-expressing regression coefficients to generate comparable results across studies regardless of data transformation have recently been developed. We examined the relative bias of three re-expression methods using simulations and 15 real data examples where the independent variable had a skewed distribution. Regression coefficients from models with log-transformed independent variables were re-expressed as though they were based on an untransformed variable. We compared the re-expressed coefficients to those from a model fit to the untransformed variable. In the simulated and real data, all three re-expression methods usually gave biased results, and the skewness of the independent variable predicted the amount of bias. Meta-analyses stratified by whether the key independent variable was log-transformed and synthesis of the stratified results without meta-analysis appear to be preferable to use of the re-expression methods examined.

Key words: Conversion, Meta-analysis as Topic, Regression Analysis, Transformation

Abbreviations:

β_{Estimand}: β coefficient from fitting $y = \beta_{\text{Estimand}} \cdot x + e$ with an ordinary least squares (OLS) model

β_{RB}: Estimator evaluated using the Rodriguez-Barranco method

β_{Dz}: Estimator evaluated using the Dzierlenga method

β_{Alt}: Estimator evaluated using the Alternative method

ADEMP: Aims, Data generating mechanism, Estimand, Methods, and Performance measures (Simulation study methodology)

DGM: Data generating mechanism

PFOA: Perfluorooctanoic acid

1. Introduction

The results of a group of studies deemed comparable can be synthesized quantitatively using meta-analysis. To base the meta-analysis on all available data, "Results extracted from study reports may need to be converted to a consistent, or usable, format for analysis" (J. P. Higgins et al., 2019). Methods of converting data presented by authors into a format suitable for meta-analysis have been well developed for intervention studies, but for studies of exposure effects the available tools are somewhat limited.

A challenge that frequently arises in the meta-analysis of non-randomized studies of exposure effects is that the authors of the studies being synthesized have made different choices about how to represent a continuous-scale exposure in their analysis. For example, the exposure might be left in its original untransformed units, log-transformed, or categorized.

Various approaches to the problem of inconsistently-expressed effect estimates have been recommended or used in practice (Campbell et al., 2020; McCabe et al., 2021; Negri et al., 2017; Ye et al., 2019). Obtaining the raw data or asking authors to re-analyze their data are the ideal solutions, though not always practical. When these options are not feasible, the results from studies using the less-frequent approach have been be excluded from the meta-analysis (McCabe et al., 2021), or preferably the results of studies that used transformed and original units are analyzed separately (Negri et al., 2017; Ye et al., 2019), and then synthesized without meta-analysis (SWiM) (Campbell et al., 2020). Some authors, however, have recently used re-expression methods to address the problem (Dzierlenga et al., 2020; Steenland et al., 2018). The validity of these re-expression methods, however, have not been evaluated in detail. Here we consider methods of re-expressing regression coefficients from linear models fit to a log-transformed exposure variable as the coefficient that would have been obtained had

the authors left the exposure in its original units. We refer to this process as re-expression of β to an untransformed basis.

An algebraic method of re-expressing regression coefficients was recently described and evaluated using one simulated data set and one set of parameters (Rodríguez-Barranco et al., 2017). Rodriguez-Barranco et al. found that in the setting of a log-transformed lognormally distributed independent variable, when the beta coefficient from a model fit to the transformed data was re-expressed to what they would have gotten had the model been fit to the untransformed data, the re-expressed coefficient was half the size of the true (fitted) coefficient. They recommended caution in applying their method when the distribution of the independent variable was markedly asymmetric. More recently, authors have developed computational methods of re-expressing coefficients from models fit to a log-transformed independent variable to approximate what would have been obtained if the model had been fit with the original unit continuous independent variable (Dzierlenga et al., 2020; Steenland et al., 2018). The basic principle is to minimize the difference between the y predicted from $y = \beta \cdot \log(x)$ and the y predicted from a $y = \beta \cdot x$ (over the same range of x) by varying β in the second equation. Figure 1 may aid visualization of the task, where y from $y = \beta \cdot \log(x)$ is shown with a light blue dotted line, and the difference in y from a straight line is minimized over a range of x. When Steenland et al. originally described the procedure it was for a fixed range of x, applicable to a specific exposure, and the validity of their method was not evaluated. Dzierlenga et al. (2020) used the same basic principle as Steenland et al. with a modification of the method to be more flexible with respect to the range of the exposure variable and found that it performed well when evaluated using data from five studies of one exposure. In addition to the above re-expression methods, we developed a third ("Alternative") estimator that is algebraic but different than that of Rodriguez-Barranco et al., and introduce it below, in the methods section.

The goal of the present project was to evaluate the validity of re-expression of regression coefficients to an untransformed basis for three methods using a wide variety of simulated and real data examples.

2 Methods

In this section, we present the simulation study that was used to evaluate the three estimators, and then describe the real datasets that were used to further evaluate the estimators. Our description of the simulation study follows the "ADEMP" format recommended by Morris et al. (2019), where ADEMP stands for Aims, Data generating mechanism, Estimand (target of analysis), Methods, and Performance measures (Morris et al., 2019). The methods subsection of the ADEMP gives a detailed specification of the estimators and is thus relatively long.

2.1 Description of the simulation in ADEMP format

Aim: To examine bias in the estimated regression coefficients (regression coefficient that would have been obtained had the original analysts not log-transformed exposure before fitting a regression model) calculated by three methods.

Data generating mechanism (DGM): An independent random variable x with a log normal distribution used to define the dependent variable $y = \beta_{\mathrm{DGM}} \cdot \log_b(x) + e$. The model parameters, possible values, and rationale for the chosen values are shown in Table 1. A $\beta_{\mathrm{DGM}} = 0$ was not studied in the simulations because it caused instability in the relative bias performance measure. A factorial simulation design was used with the parameter values indicated in Table 1. Specifically, every possible combination of parameter values was used, for a total of 960 simulation scenarios (each with $n_{\mathrm{sim}} = 2000$).

Estimand: The β coefficient from fitting $y = \beta_{\mathrm{Estimand}} \cdot x + e$ with an ordinary least squares (OLS) model.

Methods: The three estimators evaluated were 1) as described by Rodriguez-Barranco et al. (2017), 2) as described by Dzierlenga et al. (2020), and 3) an approach we introduce below and call the Alternative estimator. We refer to these as β_{RB}, β_{Dz}, and β_{Alt}, respectively.

An algebraic method for re-expression of β to an untransformed exposure was first presented by Rodriguez-Barranco et al. (2017). Equation 1 below shows their formula (See Model B in Table 1 of their publication):

$$\beta_{RB} = log_b\left(1 + \frac{c}{E[X]}\right) * \beta_{\text{from model } y = \beta \cdot \log(x)} \quad (1)$$

In equation 1, β_{RB} is the re-expressed beta coefficient using the Rodriguez-Barranco method, b is the log base used to transform x, c is the absolute change in exposure x (c = 1 unit of exposure in the present study), E[X] is the mean of the exposure, and $\beta_{...}$ is the regression coefficient from the model using the log-transformed exposure. The same formula was applied to the confidence limits of β from the log(x) model.

A computational method for re-expression of β to an untransformed exposure was developed by Steenland et al. (2018), who described their method as

> … iteratively minimizing the squared deviation of a new linear curve from the original logarithmic one, over a scale of 0 to 10 ng/ml PFOA [perfluorooctanoic acid], typical of studies in the general population. We also minimized squared deviation of a linear upper and lower confidence limit from the original logarithmic confidence interval curves. For any given study, the iteration was conducted by minimizing the sum of squares of the difference between the candidate linear curve and the logarithmic curve reported in the literature, across 10 points, at 1, 2…. through 10 ng/ml. Iteration began with an educated guess for a candidate linear curve that would approximate the

logarithmic curves and proceeded by varying the candidate linear curve until the sum of squares of the differences were minimized.

Dzierlenga et al. (2020) used this same principle to calculate β_{Dz}, though it modified it so that the method was more flexible with respect to the range of the exposure variable. The modification used an algorithmic optimization over 6 points from the 25th to the 75th percentiles (25th, 35th…75th) of the estimated exposure distribution.

The Alternative method of algebraic re-expression that we developed for this report was based on the principle of calculating, on the untransformed scale of exposure, the increment that represented a doubling, a 2.718-fold increase, or a 10-fold increase (i.e., one log unit, with a base of 2, e, or 10). This was done by subtracting or adding 0.5 units on the log scale to the log(median exposure), back-transforming the results, and taking the difference (see Equation 2).

$$I = b^{log_b(median)+0.5} - b^{log_b(median)-0.5} \tag{2}$$

In equation 2, I is the increment used to re-express β from log to linear and b is the logarithm base. Then

$$\beta_{Alt} = \beta_{\text{from model with log(x)}}/I \tag{3}$$

The same formula was applied to the confidence limits of β from the log(x) model. R scripts/functions and data files for applying each of these three re-expression methods are available in the supplemental materials (Supplemental_Code.zip).

Performance measures: We focused on relative bias, coverage probability, and the Monte Carlo standard error of the relative bias for each estimator. An example of the formula for the mean relative bias for a given scenario is:

$$Relative\ Bias = \frac{1}{n_{sim}} \sum_{i=1}^{n_{sim}} \frac{\beta_{RB_i} - \beta_{estimand_i}}{|\beta_{estimand_i}|} \tag{4}$$

Where, e.g., β_{RB_i} refers to the beta coefficient obtained from the Rodriguez-Barranco et al. estimator for the ith repetition, $\beta_{estimand_i}$ refers to the beta coefficient obtained from the ordinary least squares estimator on the untransformed, simulated data, and n_{sim} is the number of simulations conducted. Absolute value of the $\beta_{estimand}$ is used as the denominator in order to generate the correct sign for the absolute bias when both β values are negative. $\beta_{estimand}$ is used rather than β_{DGM} to calculate the relative bias in equation 4 so that the results reflect the performance of the estimator(s) in specific datasets. An example formula for the Monte Carlo standard error of the relative bias for a given scenario is:

$$Monte\ Carlo\ SE\ of\ Relative\ Bias = \sqrt{\frac{1}{n_{sim}(n_{sim}-1)}\sum_{i=1}^{n_{sim}}\left(\frac{\beta_{RB_i}-\beta_{estimand_i}}{|\beta_{estimand_i}|} - mean\ of\ relative\ bias\right)^2} \qquad (5)$$

2.2 Evaluation of the determinants of relative bias using the simulated data

After running the simulations using the parameter values shown in Table 1, for each estimator we fit ordinary least squares models of the relative bias as a function of median, σ, b (log base), n_{obs}, and β_{DGM}, and interaction terms between sigma and these variables. A both-directions stepwise approach was taken where the multiple of the number of degrees of freedom used for the penalty (k) was set to a value ~3.84 ($p < 0.05$ in Chi-square test) and the optimal model was selected by minimization of the AIC value. Each observation in the dataset analyzed was the average result from 2000 simulations. Use of the average rather than the data for all 1,920,000 (960·2000) observations resulted in essentially the same models and produced more interpretable plots.

2.3 Evaluation of the validity of the three estimators using the real data

To further evaluate the validity of re-expression methods and guide our simulations, we sought examples for various types of outcomes (dichotomous, log-continuous, untransformed continuous) and a variety of environmental agents with exposure measured using a biomarker. Environmental exposures

measured with a biomarker frequently have skewed distributions with a long tail to the right. We first identified a series of published analyses based on data that were publicly available. Second, we identified a similar series of published analyses that did not have raw data available but that presented regression results obtained with and without log transformation of the exposure.

For the example data that involved our re-analysis of published results, we chose results that could efficiently be replicated to a reasonable degree of accuracy using the originally described methods. When the authors presented results for more than one outcome or more than one exposure in a report, in general we arbitrarily chose one result that was statistically significant for inclusion in our evaluation of validity; the exception was data from Xu et al. (2020), for which we included two results. Xu et al. (2020) showed results for two different outcomes, one continuous, and one dichotomous, that were examined in relation to the same exposure; the regression coefficients were statistically significant for both.

For each real data example, we calculated the relative bias for each of the three estimators (compared to the coefficient from the untransformed exposure), and then for the 15 examples calculated the median, quartiles, and range of relative bias values for each estimator.

In the two example datasets where the relative bias in the three estimators was largest, we explored whether the exclusion of influential observations affected the accuracy of the re-expression using β_{Dz}. In two additional examples datasets where the relative bias was typical of other studies, we also examined the effect of excluding influential points on the validity of the re-expression with β_{Dz}. Influential observations were identified with a difference in beta analysis (change in beta with each observation excluded one at a time) performed on the regression using untransformed exposure. A t-test-like statistic was used to identify the 5% of points that were unusually influential ($|DFBETAS| > 2/\sqrt{n}$) (Belsley et al., 1980).

2.4 Adjustment for bias in the estimators

The regression equations we developed to evaluate the determinants of relative bias in the simulated data (Section 2.2) were used to predict the relative bias in each estimator based on σ and other parameters, as needed. The predicted relative bias was used to estimate what the value of the estimator would have been were it not biased, e.g., $\beta_{\text{Alt,adjusted}} = \beta_{\text{Alt}}/(1 + \text{predicted relative bias of } \beta_{\text{Alt}})$. We applied this to the real datasets, to see if the adjustment resulted in an improved estimator.

3. Results

3.1 Simulations

A simplified example simulation with data generated by $y = \beta_{\text{DGM}} \cdot \log_e(x)$ and parameters $\beta_{\text{DGM}} = 1$, median = 1, $\sigma = 0.5$, $SD_e = 0$ is depicted in Figure 1. In this scenario β_{RB} slightly undershot the slope estimated from the fitted regression line, whereas the β_{Dz} and β_{Alt} estimators overshot the fitted slope, by a slightly greater magnitude.

The relative bias of β_{RB} was a function of σ and the median exposure level (Figure 2A). When x was significantly skewed (e.g., $\sigma = 0.65$) and the median was 1, the relative bias was close to zero, but with other combinations of σ and median the range of bias was substantial. The coefficients for the model of relative bias in β_{RB} are shown in Table S1.

The relative bias of β_{Dz} was primarily a function of σ; the interaction of σ and n_{obs} was statistically significant but not substantially important (Figure 2B, Table S1). When x was significantly skewed (e.g., $\sigma = 0.65$), the relative bias was approximately 0.5.

The relative bias of β_{Alt} was primarily a function of σ and log base (Figure 2C, Table S1). When log base was 10, at a given value of σ the relative bias was lower than when log base was 2 or e. When log base was 2 or e, β_{Alt} performed similarly to β_{Dz}.

As noted earlier, the models of relative bias for each estimator were fit to datasets with an n of 960, where each of the 960 observations was the average of 2000 simulations for each scenario (parameter set). When the same models were fit to all of the original data points (960·2000) the model fit statistics were essentially the same (Table S2).

Figure 2D shows the relative bias after restricting the parameter set for the simulation to best display the key properties of each estimator: all depend on σ, β_{RB} additionally depends on the median, and β_{Alt} additionally depends on the log base. The overall interpretation based on the figure was that in the simulations in general, with $\sigma > 0.45$ the estimators were substantially biased except for specific circumstances where β_{RB} did well. Another way to summarize the overall findings was by the performance measures presented in Table 2, based on the results for all simulation scenarios with $|\beta_{Estimand}| > 0$. The median Monte Carlo standard error (MCSE) across all three estimators was 0.002. Thirty percent of the 2880 simulations (960 scenarios ·3 estimators) had an MCSE > 0.005 . More than 95% of all 2880 simulations had an MCSE that was ≤ 0.02 (relatively small compared with the average relative bias). Among the < 5% with an MCSE > 0.02, the nobs was 162 and the log base was 10 in all instances. The maximum MCSEs were: β_{RB}, 0.196; β_{Dz}, 0.320; and β_{Alt}, 0.262. The coverage probabilities were substantially below 95%, reflecting how infrequently the estimators performed well. Compared with β_{RB}, the other two estimators, on average, had larger positive bias, but with higher coverage probabilities.

3.2 The real data and application of the estimators to it

We identified nine published analyses of data for which the raw data were publicly available and that met our criteria for selection (Table S3) (Abraham et al., 2020; Bulka et al., 2021; Cheang et al., 2021; Lee et al., 2020; Odebeatu et al., 2019; Pilkerton et al., 2018; Stein et al., 2016; Xu et al., 2020). The results of our re-analyses were generally the same order of magnitude as those originally published

(Table S3). The specific finding that we used in the analysis and its location in the original publication are listed in Supplementary Material Table S4, as are the median, quartiles, and mean of the exposure distributions, which were estimated in some cases as indicated by table footnotes. We identified six published analyses of data where the original authors presented regression results using exposure with and without a log-transformation (Table S5) (Apelberg et al., 2007; Chen et al., 2012; Darrow et al., 2013; Hamm et al., 2010; Steenland et al., 2018; Washino et al., 2009). Five of these were included in the assessment of validity by Dzierlenga et al. (2020) (Dzierlenga et al., 2020). Among the fifteen example studies, a variety of outcomes and exposure variables were examined, though in two thirds of the studies the exposure was a perfluoroalkyl substance (either PFOA or perfluorooctane sulfonic acid).

When β was re-expressed as if it had been fit to untransformed exposure data, the range in relative bias across all three estimators was -0.49 to 16.82 (Table 3) and the interquartile ranges in relative bias were relatively wide. In the comparison of results for specific studies across re-expression methods, the relative bias was, for most of the studies, similar across methods (Table 3). These were studies where the median exposure was > 4 units (Table S4) – as would be expected based on Figure 2D. For the Lee et al. (2020) and the two Xu et al. (2020) results (Lee et al., 2020; Xu et al., 2020), however, β_{RB} performed better than the other two methods. In these three instances, the median of the exposure variable was less than one, which was not the case for the other studies (see Supplementary Materials Table S4) and the σ was > 0.8 – which is the setting where the relative bias in β_{RB} was expected to be relatively small compared with the other estimators.

Our results for Odebeatu et al. (2019) and Pilkerton et al. (2018) were the ones with the greatest discrepancy between the re-expressed β coefficients and the β fitted to the untransformed exposure (Odebeatu et al., 2019; Pilkerton et al., 2018) (Table 3). This discrepancy suggested that there may have been observations that were influential, and that the influence was affected by whether the exposure had been log-transformed. Thus, we conducted an analysis of whether exclusion of influential points

affected the accuracy of the re-expression. For comparison, similar analyses were conducted using the data from Cheang et al. (2021) and Xu et al. (2020) (dichotomous outcome), which showed smaller relative differences between the re-expressed and fitted βs. The analyses with and without the inclusion of especially influential points in the real data sets showed that the accuracy of the re-expression estimators was affected by their exclusion (Table S6). The relative bias was affected by the influential points more so for Odebeatu et al. (2019) and Pilkerton et al., (2019) than for Cheang et al. (2021) and Xu et al. (2020), but even with the exclusion of influential observations the re-expression methods still performed poorly.

When we used the regression equations (Table S3) to predict the relative bias in the estimators when applied to each of the real data examples, and then adjusted the re-expressed β to remove the bias, the adjusted βs, on average, showed less bias, but as before, the interquartile range of the adjusted relative bias was wide (Table 4).

4. Discussion

In the simulations, the bias in each of the three estimators was evaluated in relation to the median of the exposure variable, the skewness in the exposure variable, the log base used to transform the exposure variable, the β in the model generating the data, and the number of observations simulated. For all three re-expression methods, the relative bias was more positive as the skewness of the exposure distribution increased. The relative bias in β_{RB} was also determined by the median of the exposure distribution, and the relative bias in β_{Alt} was also affected by the base of the log used to transform the exposure variable. Although a few specific circumstances were found where the relative bias in a given re-expression method was lower, in general, when the skewness of x was large enough that a log transformation might be applied, the methods gave results that were sufficiently biased that their use would not be advisable. The results from applying the re-expression methods to real datasets

generally agreed with those from the simulation, but their performance was worse than predicted based on the simulations. The poor performance in the real data was not much affected by the exclusion of influential observations. The especially poor performance of the re-expression methods in the case of the Odebeatu et al. (2019) data may have been due to the small size of the slope being re-expressed.

Rodriguez-Barranco et al. (2017) recognized the importance of skewness in causing bias in their estimator, though the degree of skewness in their simulations was not specified and only one median value was used. For the re-expression method proposed by Steenland et al. (2018), apparently it was assumed that if an exposure distribution had an upper bound near 10 units, their empirical re-expression method would be sufficiently accurate (Steenland et al., 2018). Our results suggested that the range of exposure was predictive of the validity of the re-expression only for the RB estimator. Using a re-expression method similar to Steenland et al.'s (Dzierlenga et al., 2020), Dzierlenga et al. (2020) conducted a validation study, using only five studies. Three of the studies in that validation had the least bias in the present study; their relatively small number of validation studies may have led to an overly optimistic appraisal of the method.

In this report we focused on re-expressing regression coefficients from linear models fit to a log-transformed exposure variable. We could have also addressed the opposite: re-expression of regression coefficients from linear models fit to the untransformed exposure variable. To simplify the manuscript, we did not address this opposite type of re-expression. In risk assessment, results based on untransformed exposure are of greatest use, hence our focus on expressing all results in arithmetic units.

The results of this assessment of validity have implications for systematic reviewers and meta-analysts considering or using these re-expression methods. The bias due to re-expression with the three methods evaluated was affected by the skewness of the exposure variable, and, for some estimators,

the median exposure or the type of transformation used. Meta-analyses stratified by whether the key independent variable was log-transformed and synthesis of the stratified results without meta-analysis (McKenzie & Brennan, 2022) may be preferable to use of the re-expression methods studied here.

Data Availability Statement

The data used in this manuscript are freely available from their respective publications or upon request from the original authors. R scripts/functions and data files for applying each of these three re-expression methods are available in the supplemental materials (Supplemental_Code.zip).

Funding Statement

Support for this project (Linakis, Van Landingham, Longnecker) came from 3M to Ramboll. 3M did not influence the research in any way and encouraged publication in a peer-reviewed journal.

Conflict of Interest Disclosure

3M, the company that funded this research, was not involved in the preparation of the manuscript. The authors retained sole control of the manuscript content and the findings, and statements in this paper are those of the authors and not those of the author's employer or the sponsors. No authors were directly compensated by 3M.

Ethics Approval Statement

Not applicable

Patient Consent Statement

Not applicable

Permission to Reproduce Material from Other Sources

Not applicable

Clinical Trial Registration

Not applicable

Author roles:

Dr. Linakis ran the re-expression algorithm on the examples and simulation datasets, conducted the analysis of factors that determine validity, computed results using the other re-expression methods, and contributed to the draft of the manuscript.

Ms. Van Landingham replicated the analysis of the example NHANES studies, conducted the influence analyses, and contributed to the draft of the manuscript.

Dr. Gasparini helped to critically evaluate and provide feedback on the methods and presentation of the results, and contributed to the draft of the manuscript.

Dr. Longnecker conceived of the study, designed it, obtained funding for it, and drafted the manuscript.

Acknowledgments

Dr. Michael W. Dzierlenga provided critical comments on an early draft of the manuscript.

Figure Captions

Figure 1: Plot of simulated values of y as a function of x from $y = \ln(x)$(curve), along with slopes obtained by four methods (diagonal lines). The gold solid line represents a slope (β_{Estimand}) fitted with the model $y = \alpha + \beta x$. The three dashed lines are estimates of β_{Estimand} obtained by the re-expression methods described in the text. Vertical lines indicate the first and third quartiles of the x-values. The intercepts of the diagonal lines have been adjusted to emphasize the similarity of the slopes in the interquartile range.

Figure 2: Plots of relative bias as a function of skewness (σ) in the exposure x, by type of estimator. Individual points represent the average result ($n_{\text{sim}} = 2000$) for each simulation scenario. A total of 890 of the possible 1,920,000 observations (960 scenarios x 2000 simulations) were not used in the calculation of the average results because β_{estimand} was < 0.0001 (essentially zero). Lines represent quadratic fits to the data for a specified prediction equation and set of values of independent variables (see text). Note that data have been artificially spread along the x-axis for visualization purposes, all actual x-values are the closest black vertical line (0.25, 0.45, 0.65, or 0.85). Figures A-C show points for 768 simulation scenarios ($\beta_{\text{DGM}} > 0$, see Figure S1 for plots including $\beta_{\text{DGM}} < 0$); Figure D shows points for a subset of scenarios (n = 32) chosen because they demonstrate differences among the estimator properties. A) Rodriguez-Barranco estimator, B) Dzierlenga estimator, and C) Alternative estimator. D) Shows all 3 estimators in the same plot with a subset of the data of the simulation data where $\beta_{\text{DGM}} = 0.5$, log base = 2 or 10, and median value = 0.5 or 8.

References

Abraham, K., Mielke, H., Fromme, H., Völkel, W., Menzel, J., Peiser, M., Zepp, F., Willich, S. N., & Weikert, C. (2020). Internal exposure to perfluoroalkyl substances (PFASs) and biological markers in 101 healthy 1-year-old children: Associations between levels of perfluorooctanoic acid (PFOA) and vaccine response. *Archives of Toxicology*, *94*(6), 2131–2147. https://doi.org/10.1007/s00204-020-02715-4

Apelberg, B. J., Witter, F. R., Herbstman, J. B., Calafat, A. M., Halden, R. U., Needham, L. L., & Goldman, L. R. (2007). Cord serum concentrations of perfluorooctane sulfonate (PFOS) and perfluorooctanoate (PFOA) in relation to weight and size at birth. *Environmental Health Perspectives*, *115*(11), 1670–1676. https://doi.org/10.1289/ehp.10334

Belsley, D. A., Kuh, E., & Welsch, R. E. (1980). *Regression diagnostics: Identifying influential data and sources of collinearity*. Wiley.

Bulka, C. M., Avula, V., & Fry, R. C. (2021). Associations of exposure to perfluoroalkyl substances individually and in mixtures with persistent infections: Recent findings from NHANES 1999-2016. *Environmental Pollution (Barking, Essex: 1987)*, *275*, 116619. https://doi.org/10.1016/j.envpol.2021.116619

Campbell, M., McKenzie, J. E., Sowden, A., Katikireddi, S. V., Brennan, S. E., Ellis, S., Hartmann-Boyce, J., Ryan, R., Shepperd, S., Thomas, J., Welch, V., & Thomson, H. (2020). Synthesis without meta-analysis (SWiM) in systematic reviews: Reporting guideline. *BMJ*, *368*, l6890. https://doi.org/10.1136/bmj.l6890

Cheang, I., Liao, S., Zhu, X., Lu, X., Zhu, Q., Yao, W., Zhou, Y., Zhang, H., & Li, X. (2021). Association of acrylamide hemoglobin biomarkers with serum lipid levels in general US population: NHANES 2013-2016. *Ecotoxicology and Environmental Safety*, *214*, 112111. https://doi.org/10.1016/j.ecoenv.2021.112111

Chen, M.-H., Ha, E.-H., Wen, T.-W., Su, Y.-N., Lien, G.-W., Chen, C.-Y., Chen, P.-C., & Hsieh, W.-S. (2012). Perfluorinated compounds in umbilical cord blood and adverse birth outcomes. *PloS One*, *7*(8), e42474. https://doi.org/10.1371/journal.pone.0042474

Darrow, L. A., Stein, C. R., & Steenland, K. (2013). Serum perfluorooctanoic acid and perfluorooctane sulfonate concentrations in relation to birth outcomes in the Mid-Ohio Valley, 2005-2010. *Environmental Health Perspectives*, *121*(10), 1207–1213. https://doi.org/10.1289/ehp.1206372

Dzierlenga, M. W., Crawford, L., & Longnecker, M. P. (2020). Birth weight and perfluorooctane sulfonic acid: A random-effects meta-regression analysis. *Environmental Epidemiology (Philadelphia, Pa.)*, *4*(3), e095. https://doi.org/10.1097/EE9.0000000000000095

Hamm, M. P., Cherry, N. M., Chan, E., Martin, J. W., & Burstyn, I. (2010). Maternal exposure to perfluorinated acids and fetal growth. *Journal of Exposure Science & Environmental Epidemiology*, *20*(7), 589–597. https://doi.org/10.1038/jes.2009.57

Higgins, J. P., Li, T., & Deeks, J. J. (2019). Choosing effect measures and computing estimates of effect. In *Cochrane Handbook for Systematic Reviews of Interventions* (pp. 143–176). John Wiley & Sons, Ltd. https://doi.org/10.1002/9781119536604.ch6

Lee, S., Min, J.-Y., & Min, K.-B. (2020). Female Infertility Associated with Blood Lead and Cadmium Levels. *International Journal of Environmental Research and Public Health*, *17*(5), E1794. https://doi.org/10.3390/ijerph17051794

McCabe, J. J., O'Reilly, E., Coveney, S., Collins, R., Healy, L., McManus, J., Mulcahy, R., Moynihan, B., Cassidy, T., Hsu, F., Worrall, B., Murphy, S., O'Donnell, M., & Kelly, P. J. (2021). Interleukin-6, C-reactive protein, fibrinogen, and risk of recurrence after ischaemic stroke: Systematic review and meta-analysis. *European Stroke Journal*, *6*(1), 62–71. https://doi.org/10.1177/2396987320984003

McKenzie, J. E., & Brennan, S. E. (2022). Chapter 12: Synthesizing and presenting findings using other methods. In J. P. T. Higgins, J. Thomas, J. Chandler, M. Cumpston, T. Li, M. J. Page, & V. A. Welch, *Cochrane Handbook for Systematic Reviews of Interventions version 6.3*. Cochrane. www.training.cochrane.org/handbook

Morris, T. P., White, I. R., & Crowther, M. J. (2019). Using simulation studies to evaluate statistical methods. *Statistics in Medicine*, *38*(11), 2074–2102. https://doi.org/10.1002/sim.8086

Negri, E., Metruccio, F., Guercio, V., Tosti, L., Benfenati, E., Bonzi, R., La Vecchia, C., & Moretto, A. (2017). Exposure to PFOA and PFOS and fetal growth: A critical merging of toxicological and epidemiological data. *Critical Reviews in Toxicology*, *47*(6), 482–508. https://doi.org/10.1080/10408444.2016.1271972

Odebeatu, C. C., Taylor, T., Fleming, L. E., & J Osborne, N. (2019). Phthalates and asthma in children and adults: US NHANES 2007-2012. *Environmental Science and Pollution Research International*, *26*(27), 28256–28269. https://doi.org/10.1007/s11356-019-06003-2

Pilkerton, C. S., Hobbs, G. R., Lilly, C., & Knox, S. S. (2018). Rubella immunity and serum perfluoroalkyl substances: Sex and analytic strategy. *PloS One*, *13*(9), e0203330. https://doi.org/10.1371/journal.pone.0203330

Rodríguez-Barranco, M., Tobías, A., Redondo, D., Molina-Portillo, E., & Sánchez, M. J. (2017). Standardizing effect size from linear regression models with log-transformed variables for meta-analysis. *BMC Medical Research Methodology*, *17*(1), 44. https://doi.org/10.1186/s12874-017-0322-8

Steenland, K., Barry, V., & Savitz, D. (2018). Serum Perfluorooctanoic Acid and Birthweight: An Updated Meta-analysis With Bias Analysis. *Epidemiology (Cambridge, Mass.)*, *29*(6), 765–776. https://doi.org/10.1097/EDE.0000000000000903

Stein, C. R., McGovern, K. J., Pajak, A. M., Maglione, P. J., & Wolff, M. S. (2016). Perfluoroalkyl and polyfluoroalkyl substances and indicators of immune function in children aged 12-19 y: National Health and Nutrition Examination Survey. *Pediatric Research*, *79*(2), 348–357. https://doi.org/10.1038/pr.2015.213

Washino, N., Saijo, Y., Sasaki, S., Kato, S., Ban, S., Konishi, K., Ito, R., Nakata, A., Iwasaki, Y., Saito, K., Nakazawa, H., & Kishi, R. (2009). Correlations between prenatal exposure to perfluorinated chemicals and reduced fetal growth. *Environmental Health Perspectives*, *117*(4), 660–667. https://doi.org/10.1289/ehp.11681

Xu, C., Liang, J., Xu, S., Liu, Q., Xu, J., & Gu, A. (2020). Increased serum levels of aldehydes are associated with cardiovascular disease and cardiovascular risk factors in adults. *Journal of Hazardous Materials*, *400*, 123134. https://doi.org/10.1016/j.jhazmat.2020.123134

Ye, X., Kong, W., Zafar, M. I., & Chen, L.-L. (2019). Serum triglycerides as a risk factor for cardiovascular diseases in type 2 diabetes mellitus: A systematic review and meta-analysis of prospective studies. *Cardiovascular Diabetology*, *18*(1), 48. https://doi.org/10.1186/s12933-019-0851-z

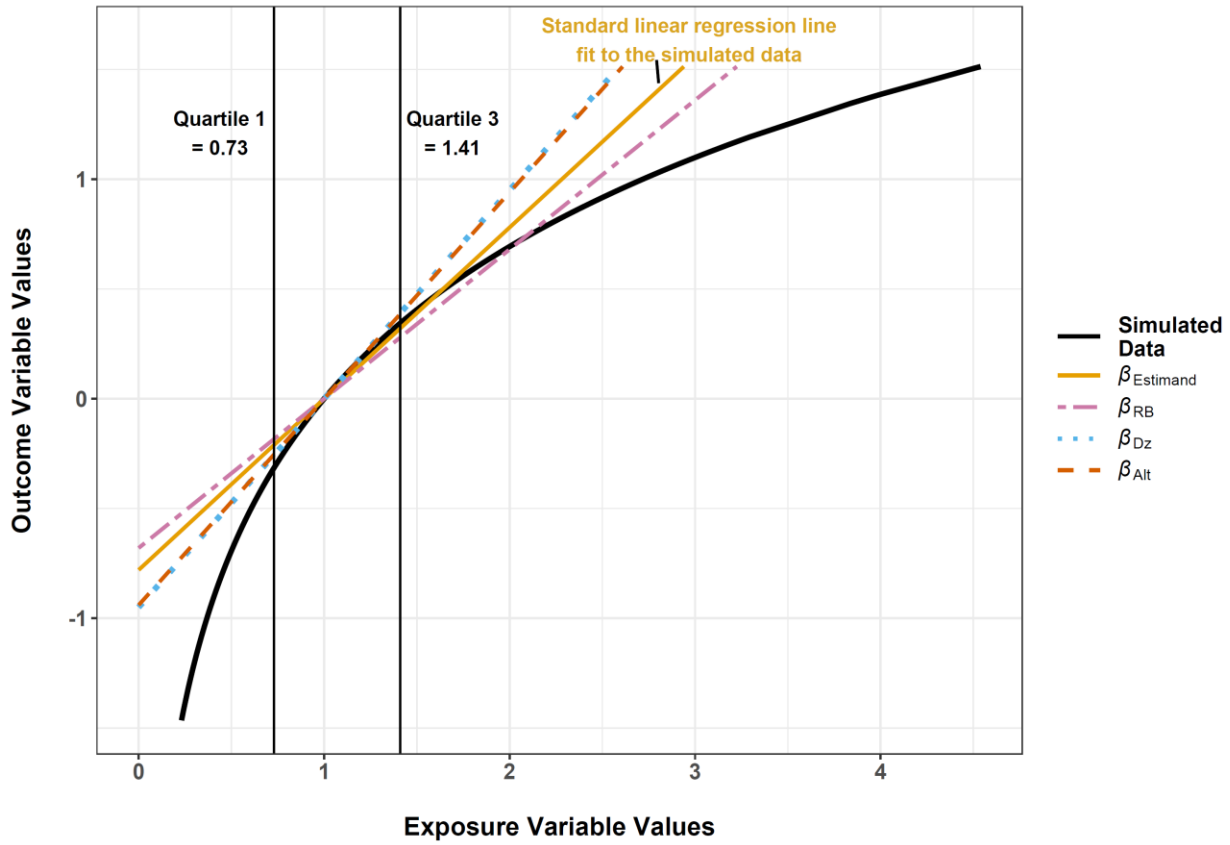
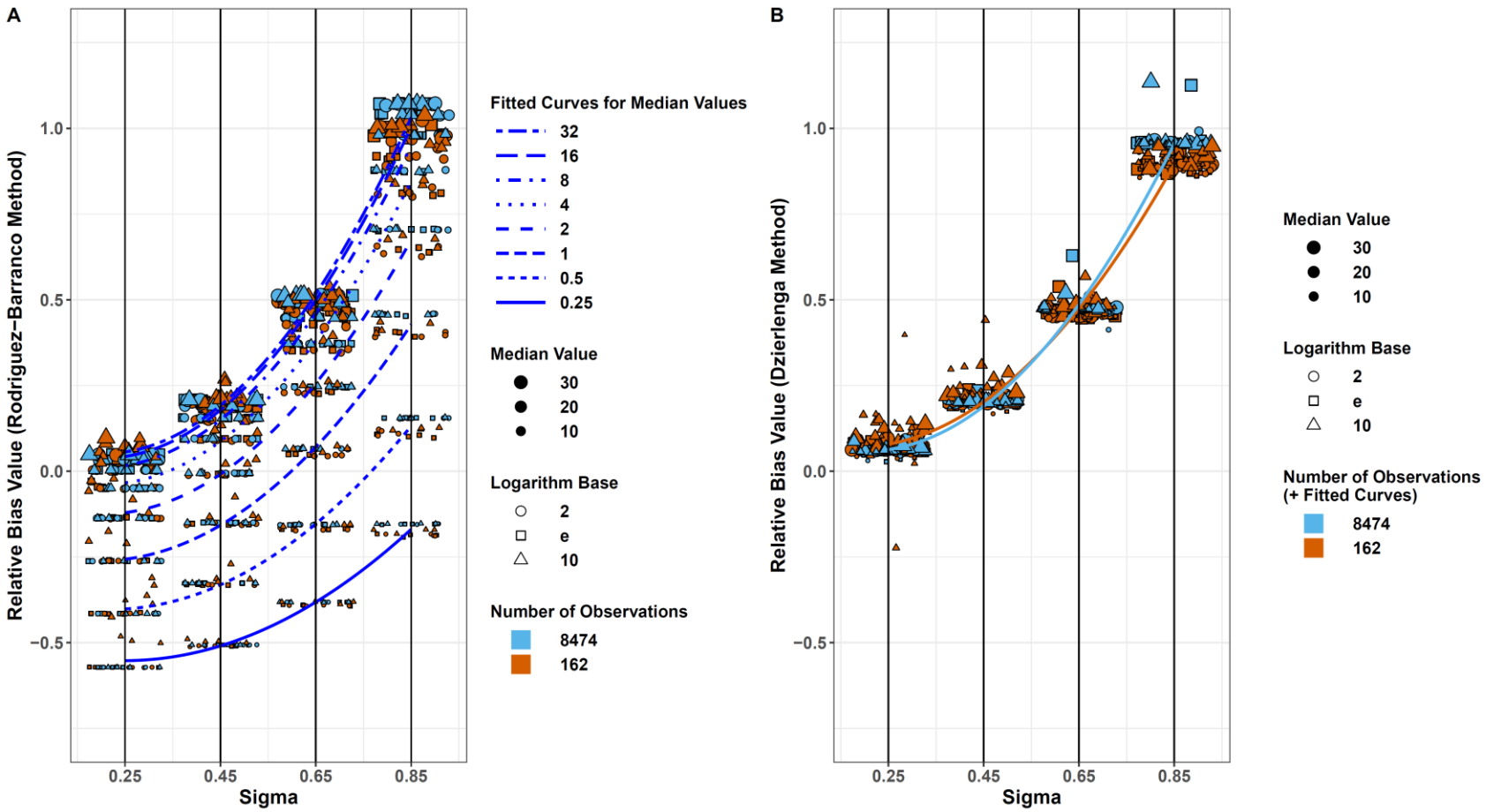


Figure 1: Plot of simulated values of *y* as a function of *x* from *y* = ln(*x*)(curve), along with slopes obtained by four methods (diagonal lines). The gold solid line represents a slope ($\beta_{Estimand}$) fitted with the model y = α + βx. The three dashed lines are estimates of $\beta_{Estimand}$ obtained by the re-expression methods described in the text. Vertical lines indicate the first and third quartiles of the x-values. The intercepts of the diagonal lines have been adjusted to emphasize the similarity of the slopes in the interquartile range.


A
Relative Bias Value (Rodriguez−Barranco Method)
1.0
0.5
0.0
−0.5
0.25
0.45
0.65
0.85
Sigma
Fitted Curves for Median Values
32
16
8
4
2
1
0.5
0.25
Median Value
30
20
10
Logarithm Base
2
e
10
Number of Observations
8474
162
B
Relative Bias Value (Dzierlenga Method)
1.0
0.5
0.0
−0.5
0.25
0.45
0.65
0.85
Sigma
Median Value
30
20
10
Logarithm Base
2
e
10
Number of Observations
(+ Fitted Curves)
8474
162

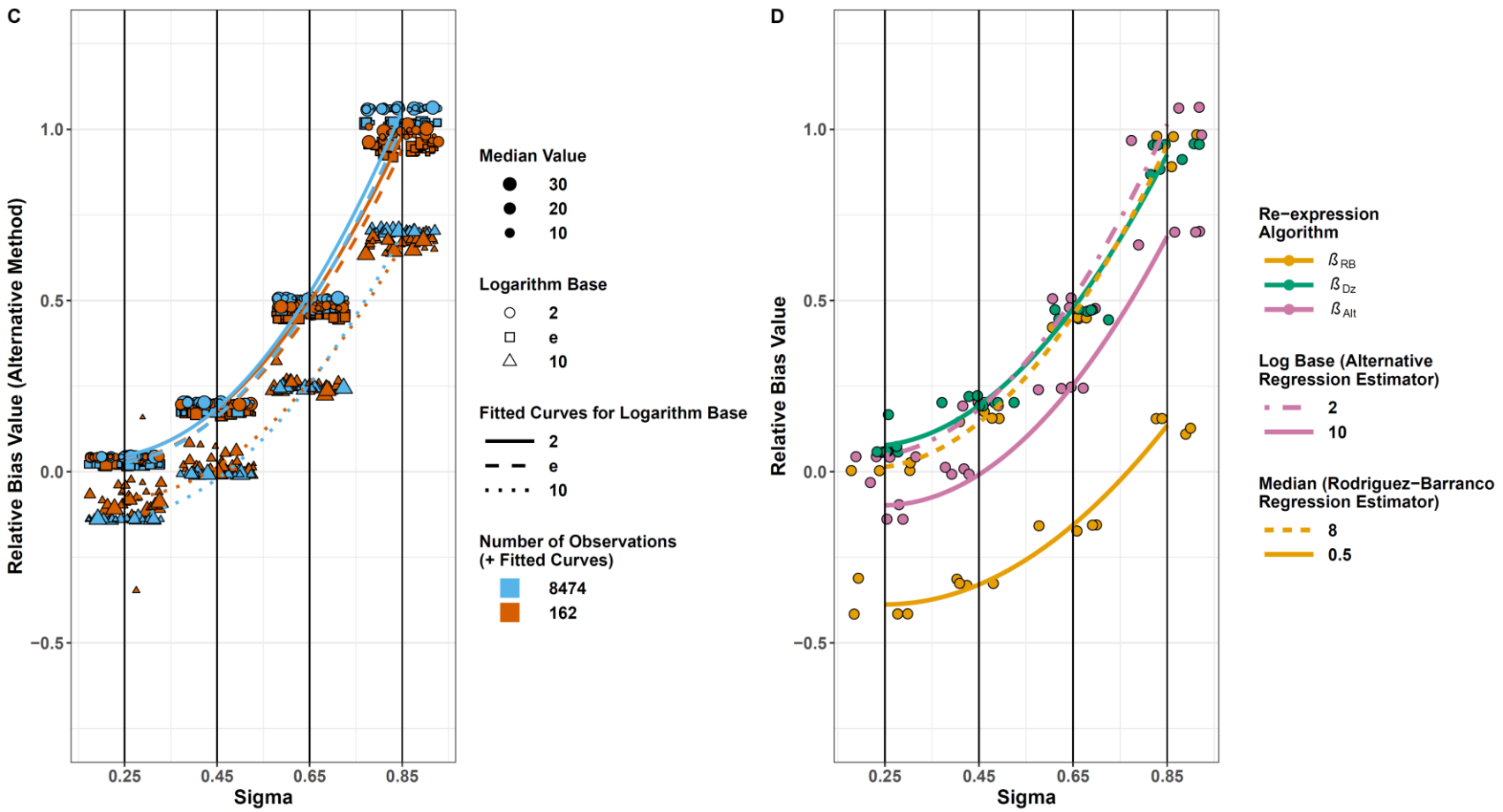


Figure 2: Plots of relative bias as a function of skewness (σ) in the exposure x, by type of estimator. Individual points represent the average result ($n_{sim} = 2000$) for each simulation scenario. A total of 890 of the possible 1,920,000 observations (960 scenarios x 2000 simulations) were not used in the calculation of the average results because $\beta_{estimand}$ was < 0.0001 (essentially zero). Lines represent quadratic fits to the data for a specified prediction equation and set of values of independent variables (see text). Note that data have been artificially spread along the x-axis for visualization purposes, all actual x-values are the closest black vertical line (0.25, 0.45, 0.65, or 0.85). Figures A-C show points for 768 simulation scenarios ($\beta_{DGM} > 0$, see Figure S1 for plots including $\beta_{DGM} < 0$); Figure D shows points for a subset of scenarios (n = 32) chosen because they demonstrate differences among the estimator properties. A) Rodriguez-Barranco estimator, B) Dzierlenga estimator, and C) Alternative

estimator. D) Shows all 3 estimators in the same plot with a subset of the data of the simulation data where $\theta_{DGM} = 0.5$, log base = 2 or 10, and median value = 0.5 or 8.

Table 1. Values of parameters used in the simulation and rationale for their choice. Number of simulations = 2000.

Parameter	Possible values	Rationale for the choice
n_{obs}	162, 8474	Quantiles (10th, 90th) of the distribution of sample sizes for the 15 real data examples.
e	Selected from a normal distribution with mean = 0 and SD = SD_e*	Standard deviation selected to result in an R^2 ~ 0.2, to make realistic models
β_{DGM}	-15, 0.5, 1, 10, 30	Broad range of effect sizes encompassing those in the real data examples.
Logbase	2, e, 10	Log bases used in the 15 real data examples.
σ	0.25, 0.45, 0.65, 0.85	Selected values cover the approximate range of σ values in the 15 real data examples.
median (μ = log(median))	0.25, 0.5, 1, 2, 4, 8, 16, 32	Selected values cover the range of median exposure in the 15 real data examples.

*SD_e was calculated using the following method: $SD_e = |\beta_{DGM} * \log(\text{median} + 2) * Mult_{SD}|$ where β_{DGM} and median are from the table above and $Mult_{SD}$ is a multiplier to prevent extremely large variability. For each median, the $Mult_{SD}$ was as follows: median 0.25, $Mult_{SD}$ 1; 0.5, 0.886; 1, 0.771; 2, 0.657; 4, 0.543; 8, 0.429; 16, 0.315; 32, 0.120). $Mult_{SD}$ was selected based on values that will introduce enough error to result in r^2 values around 0.2 (similar to many epidemiological studies) for estimation of $\beta_{Estimand}$.

Table 2: Performance measures based on simulations with all possible values for each parameter for a total of 960 simulation scenarios with 2000 simulations per scenario (n = 1,919,110).[a]

Estimator	Average absolute relative bias	IQR of absolute relative bias	Average coverage probability (%)[b]
β_{RB}	0.362	0.120-0.499	36.5
β_{Dz}	0.430	0.112-0.597	39.4
β_{Alt}	0.389	0.090-0.558	45.5

[a] A total of 890 of the possible 1,920,000 observations were not used in the calculations because β_{estimand} was < 0.0001 (essentially zero).

[b]Calculated as the sum of the replicates where the re-expression method confidence interval included the observed β (β_{estimand}) divided by the total number of 1,919,110 replicates.

Table 3: Comparison of fitted and re-expressed β coefficients and relative bias in β for three methods of re-expression

First author, year	$\beta_{Estimand}$ from analysis of raw data	β_{RB}[a]	Relative Bias[b]	β_{Dz}	Relative Bias[b]	β_{Alt}[c]	Relative Bias[b]
Abraham, 2020	-0.0636 $\log_e$/ ng·ml^{-1}	-0.0324	-0.49	-0.0368	-0.42	-0.0384	-0.40
Apelberg, 2007	-12.9 g/(ng/ml)	-11.7	-0.10	-13.2	0.03	-13.2	0.03
Bulka, 2021	0.015 $\log_e$(OR)/ ng·ml^{-1}	0.043	1.87	0.0512	2.41	0.0531	2.54
Cheang, 2021	0.192 mg·dl/pmol·g^{-1}	0.335	0.74	0.352	0.83	0.353	0.84
Chen, 2012	-11.3 g/(ng/ml)	-16.3	0.44	-17.9	0.58	-17.8	0.58
Darrow, 2013	-2.3 g/(ng/ml)	-1.95	-0.15	-2.02	-0.12	-2.00	-0.13
Hamm, 2010	1.5 g/(ng/ml)	3.66	1.44	3.92	1.62	3.85	1.57
Lee, 2020	1.444 $\log_e$(OR)/µg·dl^{-1}	1.31	-0.09	3.30	1.28	3.46	1.4
Odebeatu, 2019	6.86x10^{-4} $\log_e$(OR)/ng·ml^{-1}	0.0114	15.69	0.0122	16.82	0.0116	15.86
Pilkerton, 2018	-0.0049 $\log_e$/ng·ml^{-1}	-0.0207	3.22	-0.0302	5.17	-0.0306	5.24
Steenland, 2009	0.00105 $\log_e$/ ng·ml^{-1}	0.00116	0.11	0.00127	0.21	0.00126	0.20
Stein, 2016	-0.0039 $\log_e$/ng·ml^{-1}	-0.00676	0.73	-0.0069	0.77	-0.00694	0.78
Washino, 2009	-10.94 g/(ng/ml)	-11.4	0.04	-12.1	0.10	-10.0	-0.08
Xu, 2020a	0.513 $\log_e$(OR)/ ng·ml^{-1}	0.508	-0.01	0.962	0.87	1.01	0.96
Xu, 2020b	29.3 mg·dl/ng·ml^{-1}	25.4	-0.13	48.1	0.64	50.4	0.72
Median			0.11		0.77		0.78
25th percentile			-0.095		0.155		0.115
75th percentile			1.09		1.45		1.485
Minimum			-0.49		-0.42		-0.40
Maximum			15.69		16.82		15.86

[a] Using the notation of Rodriguez-Barranco et al., k = base of log transformation used; c = 1.

[b] Proportional difference between β in column to the left compared with the one from the analysis of raw data, calculated using the same method as in previous table ((beta in column to left – beta from analysis of raw data)/beta from analysis of raw data). Note that the β in column to the left was calculated using β with different units in denominator than for the raw data analysis shown in the table.

[c] Let the log unit increment I in untransformed units = $b^{(\log_b(\text{median}) + 0.5)} - b^{(\log_b(\text{median}) - 0.5)}$, where b = 2, e, or 10, depending on the base. For observed β_o with units of $\Delta y/\Delta \log_b(x)$, to get re-expressed β_r with units $\Delta y/\Delta x$, calculate $\beta_r = \beta_o/I$. If the units of β_o are $\Delta y/\Delta x$, to get β_r with units $\Delta y/\Delta \log_b(x)$, calculate $\beta_r = \beta_o \cdot I$.

Table 4: Comparison of fitted and re-expressed β coefficients and proportional difference in β for three methods of re-expression, adjusted by their OLS describing the relationship between sigma and relative bias.

First author, year	$\beta_{Estimand}$ from analysis of raw data	β_{RB}[a]	Relative Bias[b]	β_{Dz}	Relative Bias[b]	β_{Alt}[c]	Relative Bias[b]
Abraham, 2020	-0.0636 $\log_e$/ ng·ml^{-1}	-0.0202	-0.68	-0.0214	-0.66	-0.0214	-0.66
Apelberg, 2007	-12.9 g/(ng/ml)	-10.6	-0.18	-9.41	-0.27	-9.37	-0.27
Bulka, 2021	0.015 $\log_e$(OR)/ ng·ml^{-1}	0.0339	1.26	0.0303	1.02	0.0301	1.01
Cheang, 2021	0.192 mg·dl/pmol·g^{-1}	0.224	0.17	0.288	0.50	0.290	0.51
Chen, 2012	-11.3 g/(ng/ml)	-14.8	0.31	-12.8	0.14	-12.8	0.13
Darrow, 2013	-2.3 g/(ng/ml)	-1.69	-0.27	-1.56	-0.32	-1.56	-0.32
Hamm, 2010	1.5 g/(ng/ml)	3.79	1.52	3.24	1.16	3.24	1.16
Lee, 2020	1.444 $\log_e$(OR)/μg·dl^{-1}	0.957	-0.34	1.79	0.24	1.79	0.24
Odebeatu, 2019	6.86x10^{-4} $\log_e$(OR)/ng·ml^{-1}	0.00374	4.45	0.00361	4.26	0.00364	4.30
Pilkerton, 2018	-0.0049 $\log_e$/ng·ml^{-1}	-0.021	3.26	-0.0232	3.74	-0.0231	3.71
Steenland, 2009	0.00105 $\log_e$/ ng·ml^{-1}	0.00088	-0.16	0.00090	-0.14	0.00090	-0.14
Stein, 2016	-0.0039 $\log_e$/ng·ml^{-1}	-0.00527	0.35	-0.00542	0.39	-0.00542	0.39
Washino, 2009	-10.94 g/(ng/ml)	-11.6	0.06	-9.34	-0.15	-9.40	-0.14
Xu, 2020a	0.513 $\log_e$(OR)/ ng·ml^{-1}	0.380	-0.26	0.535	0.04	0.534	0.04
Xu, 2020b	29.3 mg·dl/ng·ml^{-1}	19.0	-0.35	26.7	-0.09	26.7	-0.09
Median			0.06		0.14		0.13
25th percentile			-0.265		-0.145		-0.140
75th percentile			0.805		0.760		0.760
Minimum			-0.68		-0.66		-0.66
Maximum			4.45		4.26		4.30

[a] Using the notation of Rodriguez-Barranco et al., k = base of log transformation used; c = 1.

[b] Proportional difference between β in column to the left compared with the one from the analysis of raw data, calculated using the same method as in previous table ((beta in column to left – beta from analysis of raw data)/beta from analysis of raw data). Note that the β in column to the left was calculated using β with different units in denominator than for the raw data analysis shown in the table.

[c] Let the log unit increment I in untransformed units = $b^{(\log_b(\text{median}) + 0.5)} - b^{(\log_b(\text{median}) - 0.5)}$, where b = 2, e, or 10, depending on the base. For observed β_o with units of $\Delta y/\Delta \log_b(x)$, to get re-expressed β_r with units $\Delta y/\Delta x$, calculate $\beta_r = \beta_o/I$. If the units of β_o are $\Delta y/\Delta x$, to get β_r with units $\Delta y/\Delta \log_b(x)$, calculate $\beta_r = \beta_o \cdot I$.