STATISTICS IN MEDICINE
Statist. Med. 2007; **26**:443–457
Published online 8 March 2006 in Wiley InterScience (www.interscience.wiley.com). DOI: 10.1002/sim.2538

Revisiting the relation between change and initial value: A review and evaluation

Yu-Kang Tu[1,2,*,†] and Mark S. Gilthorpe[1,‡]

[1] *Biostatistics Unit, Centre for Epidemiology and Biostatistics, LIGHT, University of Leeds, Leeds, U.K.*
[2] *Leeds Dental Institute, University of Leeds, Leeds, U.K.*

SUMMARY

The relation between initial disease status and subsequent change following treatment has attracted great interest in clinical research. However, statisticians have repeatedly warned against correlating/regressing change with baseline due to two methodological concerns known as mathematical coupling and regression to the mean. Oldham's method and Blomqvist's formula are the two most often adopted methods to rectify these problems. The aims of this article are to review briefly the proposed solutions in the statistical and psychological literature, and to clarify the popular misconception that Blomqvist's formula is superior to Oldham's method. We argue that this misconception is due to a failure to recognize that the heterogeneity of individual responses to treatment is a source of regression to the mean in the analysis of the relation between change and initial value. Furthermore, we demonstrate how each method actually answers different research questions, and how confusion arises when this is not always understood. Copyright © 2006 John Wiley & Sons, Ltd.

KEY WORDS: analysis of change; mathematical coupling; regression to the mean; Blomqvist's formula; Oldham's method

1. INTRODUCTION

The relation between initial disease status and subsequent change following treatment has attracted long-lasting interest in clinical research. In randomized controlled trials, the main research question is usually whether or not the observed change in disease status, assessed by observed change in overall means of the health outcome before and after the treatment, can be attributed to the treatment. In clinical practice, when treatments are proven to be effective, differential baseline effects are also of interest to many clinicians because sometimes they seek to identify subgroups of patients who might benefit more from one treatment than another.

[*] Correspondence to: Yu-Kang Tu, Biostatistics Unit, Centre for Epidemiology and Biostatistics, Leeds Institute of Genetics, Health, and Therapeutics, University of Leeds, 30/32 Hyde Terrace, Leeds, LS2 9LN, U.K.
[†] E-mail: y.k.tu@leeds.ac.uk, yukangtu@hotmail.com
[‡] E-mail: m.s.gilthorpe@leeds.ac.uk

Received 28 July 2005
Accepted 20 January 2006

For instance, suppose that two treatments—A and B—show similar mean treatment effects, but there is differential baseline effect in patients given treatment A and not in those given treatment B. Clinicians might decide to give treatment A to patients who suffer more serious diseases, especially if complications and/or costs of A and B differ.

Many clinical studies in fact show that patients with greater disease severity at baseline respond better to treatment [1, 2]. The relation between baseline disease severity and treatment effect has a generic name in the statistical literature: *the relation between change and initial value* [3], because treatment effect is evaluated by measuring the change of variables from their initial (baseline) values. In psychology, it is also well-known as *the law of initial value* [4]. However, testing the relation between change and initial value using correlation or regression has long been criticized by many statisticians as problematic. Two methodological concerns known as *mathematical coupling* [5–10] and *regression to the mean* [11–17] have been raised as the causes of the problem in testing the relation between change and initial value.

Mathematical coupling occurs when one variable directly or indirectly contains the whole or part of another, and the two variables are then analysed using correlation or regression [5]. As a result, the statistical procedure of testing the null hypothesis—that the coefficient of correlation or the slope of regression is zero—might no longer be appropriate [6], and the results need to be interpreted cautiously [5–10]. Regression to the mean occurs with any variable that fluctuates within an individual *or a population* (the latter is sometimes overlooked, as we will point out), either due to measurement error and/or physiological variation [18–22]. For instance, one is likely to obtain different readings of systolic blood pressure for the same individual when a series of measurements are made over a short time period. This can be attributed to either the 'true' underlining blood pressure fluctuating around a mean value (i.e. assuming blood pressure can be taken without measurement error), or the device used to measure blood pressure (or the person who uses the device) is not entirely reliable (this is treated as measurement error), or both. Campbell and Kenny pointed out that any factor that makes the correlation between two variables less than perfect can cause regression to the mean [23].

Several alternative statistical methods have been proposed in the medical and statistical literature to overcome problems in testing the relation between change and initial value using correlation or regression. The aim of this article is to clarify a widespread conceptual confusion around regression to the mean within the statistical literature and to correct a popular misconception about the 'correct' analysis of the relationship between change and initial value in certain scenarios. We review the proposed solutions to testing the relation between change and initial value in the statistical and psychological literature and show that, although the problem has been known for a long time, current recommendations are inadequate.

2. WHY SHOULD CHANGE NOT BE REGRESSED ON INITIAL VALUE? A REVIEW OF THE PROBLEM

Although many articles and textbooks of medical statistics warn against correlating or regressing change on initial value, it is far from clear what the problem is with this practice. The most commonly given reason is that testing the relation between change and initial value using correlation or regression suffers *regression to the mean*. Obviously, this simple explanation merits a further query: why does testing the relation between change and initial value using correlation or regression suffer regression to the mean? The most common answer given in the literature seems to be that regression to the mean is caused by biological variation and/or measurement error in the assessment of initial values. An explanation following this line of reasoning can be found in an article by Healy [24], supposing that the true (unobserved) initial value is X and the true change is D, so the true follow-up value is $Y=X-D$. The observed initial value is then $x=X+e_X$ and the observed follow-up value $y=Y+e_Y$, so the observed change $d=x-y=D+e_X-e_Y$. Therefore, testing the relation between change and initial value is to test the relation between x and d. Since e_X occurs in both x and d, their relation is likely to be positive. If change is defined as $y-x$ (as in psychology), the relation between change and initial value will tend to be negative. As this error term occurs in both change and initial value, testing the relation between change and initial value using correlation or regression is biased. As a result, the problem of regression to the mean in testing the relation between change and initial value seems to be caused only by measurement error in the initial value. Thus, the whole problem in testing the relation between change and initial value seems to be reduced to the problem of measurement error, and it is obvious why any statistical method that purports to correct the bias caused by measurement error in the initial value might seem to provide a solution.

3. PROPOSED SOLUTIONS IN THE LITERATURE

3.1. Blomqvist's formula

Blomqvist in 1977 [3] devised a formula to correct for measurement errors in initial values, to obtain an unbiased estimate of regression slopes in analysing change and initial value. Blomqvist's formula is given as [17]

$$b_{\text{true}}=\frac{b_{\text{observed}}-k}{1-k} \tag{1}$$

where b_{true} is the true regression slope, b_{observed} is the observed regression slope, k the ratio of the measurement error variance for x and the observed variance of x. If b_{true} is close to zero, it is then assumed that there is no evidence that the treatment effect is dependent upon baseline disease severity. Blomqvist's formula corrects for the bias caused by measurement error and/or biological variation in the initial values. To use this formula requires an independent (external) estimate of the error variance, which is often obtained by measuring initial values repeatedly in a short interval before the intervention is administered.

3.2. Oldham's method: testing change and average

In a seminal paper published in 1962, Oldham [12] warned against testing the relation between treatment effect of anti-hypertensive therapy and patients' initial blood pressure. One of his arguments, which has subsequently been used repeatedly by other studies [6, 16], is that for two series of independent random numbers x and y with the same standard deviation, one observes a strong correlation ($1/\sqrt{2}\approx 0.71$) between $x-y$ and x. Following previous notation, let x be the pre-treatment (initial) value and y the post-treatment value. The Pearson correlation between change $(x - y)$ and pre-treatment value (x) is [12]

$$\mathrm{Corr}[x - y, x] = r_{x-y,x} = \frac{s_x - r_{xy}s_y}{\sqrt{s_x^2 + s_y^2 - 2r_{xy}s_xs_y}} \tag{2}$$

where s_x^2 is the variance of the x, s_y^2 is the variance of y, and r_{xy} is the correlation between x and y.

If s_x^2 and s_y^2 are equal, equation (2) reduces to $r_{x-y,x} = \sqrt{(1 - r_{xy})/2}$. This formula shows that unless r_{xy} is unity, $r_{x-y,x}$ will never be 0. When r_{xy} is less than 1, the correlation between baseline and change, $r_{x-y,x}$, will always be positive; a very likely situation when repeated measurements are made on the same individuals. When r_{xy} is close to zero, i.e. there is poor correlation between pre- and post-treatment values, the positive association between baseline and change will be large. As both x and y are measured with error, r_{xy} will be always less than 1 and $r_{x-y,x}$ will be positive.

The solution proposed by Oldham did not deal with the problem of measurement error in initial values directly. Oldham suggested that testing the hypothesis that treatment effects are related to baseline values should be carried out by plotting the change against the average of the pre- and post-test values, and not against the baseline values. For instance, if pre-treatment blood pressure (BP) is denoted as x and post-treatment BP as y, BP reduction after patients are given anti-hypertensive medication will be $x - y$ and the average BP will be $(x + y)/2$. To address whether or not greater baseline BP is related to greater BP reduction following treatment, Oldham's method tests the correlation between $x - y$ and $(x + y)/2$ instead of testing the correlation between $x - y$ and x. The Pearson correlation between change and average is [12]

$$\mathrm{Corr}[x - y, (x + y)/2] = \frac{s_x^2 - s_y^2}{\sqrt{(s_x^2 + s_y^2)^2 - 4r_{xy}^2 s_x^2 s_y^2}} \tag{3}$$

where s_x^2 is the variance of the x, s_y^2 is the variance of y, and r_{xy} is the correlation between x and y.

The numerator in equation (3) indicates that Oldham's method is a test of the differences in the variances between two repeated measurements, where the two variances may be correlated [12]. If there is no difference in the variances of pre-treatment BP (x) and post-treatment BP (y), the correlation using Oldham's method will be zero, i.e. the treatment effect (BP reduction) does not depend upon baseline BP. The rationale behind Oldham's method is that if, on average, greater BP reduction can be obtained for greater baseline BP, the post-treatment BP values will become 'closer' to each other, i.e. the variance of post-treatment BP (s_y^2) will *shrink* and become smaller than that of pre-treatment BP (s_x^2). In other words, if there is a *differential* treatment effect (i.e. a greater or smaller treatment effect can be achieved in subjects with greater disease severity), this will manifest as a change of variances between the two measurements. As a result, if there is no difference in the variances between baseline and post-treatment values, there is little evidence for a *differential* treatment effect across the levels of baseline values. Oldham's strategy has been proposed previously, as early as 1939 by Morgan and Pitman [25, 26] to test the equivalence of two variances, and later in 1985 by Bland and Altman [27] to compare two methods of measurement.

Whilst not everyone has agreed with Oldham on his solution [28, 29], almost everyone has agreed that it is problematic to test $x-y$ and x. In later correspondences [28, 29], it is clear that for some it is hard to understand why we should test the relation between $x-y$ and $x+y$ if our research question is to know the relation between $x-y$ and x.

3.3. *Geometrical presentation of Oldham's method*

First used by Fisher in deriving the statistical distribution of the Pearson correlation coefficient [30], vector geometry is a useful tool to provide insights to the problem of testing the relation between change and initial value. We represent the original $x=\{X_1, X_2, X_3, \ldots, X_n\}$ and $y=\{Y_1, Y_2, Y_3, \ldots, Y_n\}$ as scaled vectors $\mathbf{x}$ and $\mathbf{y}$, where each vector element is a transformation of the original data such that the length of the vectors $\|\mathbf{x}\|$ and $\|\mathbf{y}\|$ are the standard deviations (SDs) based on the original samples. Thus, $\tilde{x}_i=(x_i-\bar{x})/\sqrt{n-1}$ where $\tilde{x}_i, i=1,\ldots,n$, are the transformed vector elements of $\mathbf{x}$, and similarly for $\mathbf{y}$. Then, the correlation between variables x and y is the cosine of the angle between the vectors $\mathbf{x}$ and $\mathbf{y}$. When x and y are two independent random variables (i.e. the correlation between x and y is expected to be zero), the angle between $\mathbf{x}$ and $\mathbf{y}$ is $\pi/2$ radians (or 90°), and these two vectors are therefore orthogonal: $\mathbf{x}\perp\mathbf{y}$. When the correlation between x and y is positive, the angle between $\mathbf{x}$ and $\mathbf{y}$ will be less than $\pi/2$; when the correlation is negative, the angle will be greater than $\pi/2$. If x and y have also the same standard deviation, i.e. vectors $\mathbf{x}$ and $\mathbf{y}$ have the same length (i.e. $\|\mathbf{x}\|=\|\mathbf{y}\|$), the length of $\mathbf{x}-\mathbf{y}$ (i.e. the SD of $x+y$) will be $\sqrt{2}\|\mathbf{x}\|$, and the angle between $\mathbf{x}-\mathbf{y}$ and $\mathbf{x}$ will be $\pi/4$ radians. From elementary trigonometry: $\cos(\pi/4)=1/\sqrt{2}\approx 0.71$.

Applying vector geometry to Oldham's method, it becomes apparent that $\mathbf{x}+\mathbf{y}$ and $\mathbf{x}-\mathbf{y}$ are orthogonal vectors *if and only if* $\mathbf{x}$ and $\mathbf{y}$ have equal lengths. This property holds irrespective of the correlation between $\mathbf{x}$ and $\mathbf{y}$ (Figure 1). Therefore, the angle between vectors $\mathbf{x}+\mathbf{y}$ and $\mathbf{x}-\mathbf{y}$ is determined by the length of $\mathbf{x}$ and the length of $\mathbf{y}$; that is, the correlation between variables $x+y$ and $x-y$ is determined by the variances of x and y.

3.4. *Variance ratio test*

This test is very similar to Oldham's method in strategy, because it is mainly to test the equivalence of variances between two correlated variables, such as two repeated measurements. Based on the same assumptions as those for Oldham's method, the variance ratio s_x^2/s_y^2 is proposed as an appropriate test, by assessing the equality of the correlated variances [31], yielding a statistic that follows the t-distribution with $n-2$ degrees of freedom [32], and which is non-significant if the variances are similar

$$t=\frac{(s_x^2-s_y^2)\sqrt{n-2}}{2s_xs_y\sqrt{1-r_{xy}^2}} \tag{4}$$

where s_x^2, s_y^2, and r_{xy} are as defined previously. Let $d=x-y$, $s=x+y$ and r_{ds} be the correlation between d and s, then equation (4) is exactly equivalent to the one proposed by

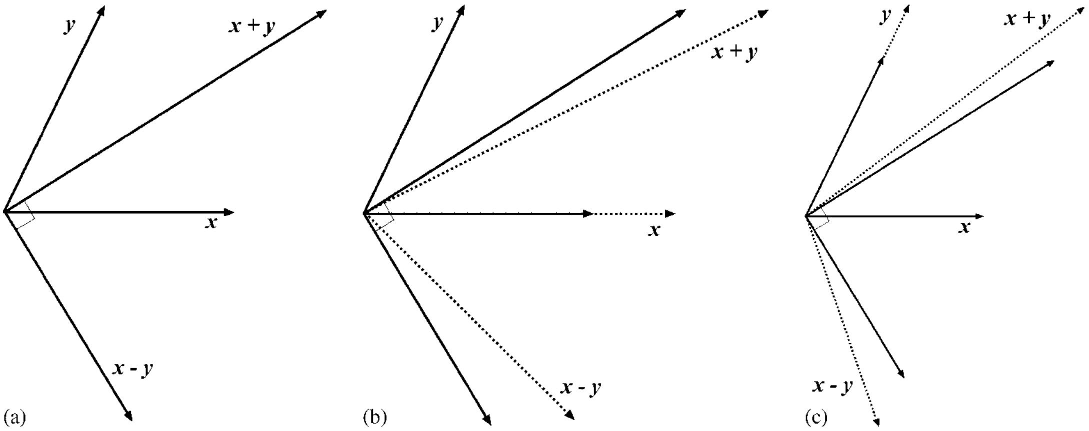

Figure 1. (a) When the two vectors **x** and **y** have equal length, the angle between vectors $\mathbf{x}+\mathbf{y}$ and $\mathbf{x}-\mathbf{y}$ will always be $\pi/2$, irrespective of the angle between **x** and **y**; (b) when the two vectors **x** and **y** have unequal length, the angle between vectors $\mathbf{x}+\mathbf{y}$ and $\mathbf{x}-\mathbf{y}$ will be no longer $\pi/2$, irrespective of the angle between **x** and **y**: when the length of **x** is greater than that of **y**, the angle between vectors $\mathbf{x}+\mathbf{y}$ and $\mathbf{x}-\mathbf{y}$ will be less than $\pi/2$; and (c) when the length of **x** is smaller than that of **y**, the angle will be greater than $\pi/2$.

Maloney and Rastogi [33]

$$t=\frac{r_{ds}\sqrt{n-2}}{\sqrt{1-r_{ds}^2}} \tag{5}$$

Like Oldham's method, this test assumes the error variances in x and y are independent and equal, and x and y follow a normal distribution. This test has been used as a two-sided test to assess whether or not there is any baseline effect [32]. It has also been used as a one-sided test when one variance is anticipated to be larger than the other [34].

3.5. *Structural regression*

Structural regression [35] has been proposed in the psychological literature to test the relation between the unobserved, true X and Y by correcting for measurement error in the observed variables x and y. If the regression slope for Y regressed on X is less than 1 (i.e. the regression slope for $X-Y$ regressed on Y is greater than zero), it indicates that the change $(X-Y)$ is dependent upon baseline value (X). In medical statistics, structural regression is also used for correction for measurement error [34]. By assuming the variances of measurement errors of X (e_X) and Y (e_Y) are equivalent, the maximum likelihood estimate of the regression slope for Y on X, $\hat{\beta}$, is [36]

$$\hat{\beta}=\frac{s_y^2-s_x^2+\sqrt{(s_y^2-s_x^2)^2+4s_{xy}^2}}{2s_{xy}} \tag{6}$$

where s_{xy} is the covariance of x and y. Denoting $\mathbf{M}$ as the covariance matrix of x and y, the eigenvalues of $\mathbf{M}$, λ, are derived by solving $\mathbf{M}-\lambda\mathbf{I}=0$ (where $\mathbf{I}$ is the identity matrix)

$$\lambda_1=\frac{s_y^2+s_x^2+\sqrt{(s_y^2+s_x^2)^2-4(s_x^2s_y^2-s_{xy}^2)}}{2} \quad \text{and}$$

$$\lambda_2=\frac{s_y^2+s_x^2-\sqrt{(s_y^2+s_x^2)^2-4(s_x^2s_y^2-s_{xy}^2)}}{2}$$

It is not difficult to prove that $\hat{\beta}$ is the slope of the first principal component of x and y [35, 36], and can be estimated by $(\lambda_1-s_x^2)/s_{xy}$ [37]

$$\frac{\lambda_1-s_x^2}{s_{xy}}=\frac{\dfrac{s_y^2+s_x^2+\sqrt{(s_y^2+s_x^2)^2-4(s_x^2s_y^2-s_{xy}^2)}}{2}-s_x^2}{s_{xy}}=\frac{s_y^2-s_x^2+\sqrt{(s_y^2-s_x^2)^2+4s_{xy}^2}}{2s_{xy}}$$

From the geometrical perspective, as long as the correlation between x and y (r_{xy}) is not zero (therefore s_{xy} is not zero), the slope of the first principal component will be unity *if and only if* the variances of x and y are equivalent. Thus, using structural regression to test the slope of the first principal component ($\hat{\beta}$) is again to test the equivalence of the variances of x and y.

3.6. *Testing the correlation between change and initial value against the correct null hypothesis*

Andersen [38] has argued that due to mathematical coupling the null hypothesis for testing the correlation between $x-y$ and x ($r_{x-y,x}$) is no longer zero. However, he did not explain how to derive a correct null hypothesis. In a short note by Bartko and Pettigrew [39], they showed that the range of $r_{x-y,x}$ is restricted by the correlation between x and y (r_{xy}), and the range of $r_{x-y,x}$ is generally not between -1 and 1. In a previous study [40], we proposed that a proper null hypothesis can be derived from equation (2) by assuming $s_x=s_y$. Therefore, the correct null hypothesis for given x and y is $\sqrt{(1-r_{xy})/2}$. To compare $r_{x-y,x}$ to $\sqrt{(1-r_{xy})/2}$, both correlation coefficients need to be transformed using Fisher's z transformation [41]. Our previous study [40] found this test in general to yield comparable results to Oldham's method.

3.7. *Multilevel modelling*

Another approach is to use multilevel modelling [42, 43]. By treating the initial and post-treatment blood pressure (BP) as the lower level and individuals as the upper level, the correlation between the variance of the intercept and the variance of the slope for the covariate Time (i.e. initial and post-treatment occasions) in the multilevel model indicates the relation between baseline disease status (intercept) and treatment effect (slope). The 2-level model is written as

$$(\mathrm{BP})_{ij}=\beta_{0ij}+\beta_{1j}\,\mathrm{Time}_{ij} \tag{7}$$

It should be noted that different parameterizations of Time will yield different results [43]. For instance, when Time is coded as 0 (initial) and 1 (post-treatment), testing the correlation between intercept and slope is equivalent to testing $r_{x-y,x}$, because the intercept variance is the variance of x and the slope variance is the variance of $x-y$; when Time is centred, such as -0.5 (initial) and 0.5 (post-treatment), testing the correlation between intercept and slope is equivalent to Oldham's method because the intercept variance is the variance of $(x+y)/2$ and the slope variance is the variance of $x-y$; when Time is coded as -1 (initial) and 0 (post-treatment), testing the correlation between intercept and slope is equivalent to testing $r_{x-y,y}$ because the intercept variance is the variance of y and the slope variance is the variance of $x-y$. An advantage of multilevel modelling over other approaches is that this method can be applied to more than two measurement occasions; details can be found in our previous study [43].

4. COMPARISON BETWEEN OLDHAM'S METHOD AND BLOMQVIST'S FORMULA

In a widely cited article by Hayes [15], Oldham's method was shown to be biased towards a negative association (however, in the original article by Hayes, change was defined as $y-x$, so the bias was positive in Hayes [15]), if: (1) the individuals have been selected on the basis of high initial values; or (2) the 'true' treatment effect differs across individuals [11, 15]. Therefore, Hayes [15] recommended Blomqvist's formula as it (allegedly) performs better than Oldham's method for these two circumstances. Whilst we agree with Hayes that Oldham's method will indeed lead to biased results in scenario (1), we assert that Hayes' position on scenario (2) is a misunderstanding, and Oldham's method in fact gives rise to correct results (whereas Blomqvist's formula does not). We consider numerical examples similar to those used by Hayes to illustrate our assertion and we also perform 10 000 simulations for each scenario. All hypothetical data generated and statistical analyses undertaken were performed using *R* (version 2.1.1. *R* development Core Team, Vienna, Austria, 2005).

4.1. Simulation

Let X be the unobserved *true* initial systolic blood pressure (SBP) values with mean 150 mmHg and standard deviation (SD) of 15 mmHg. We assume a true treatment effect D of exactly 20 mmHg across all levels of X, i.e. there is no relation between initial and change values. Therefore, the unobserved true post-treatment values Y $(=X-D)$, will have a mean of 130 mmHg and the same SD of 15 mmHg. The observed initial SBP is $x=X+e_X$, and the observed post-treatment SBP (y) is equal to $Y+e_Y$, where e_X and e_Y are measurement errors and/or biological variations for X and Y, respectively, with zero means and equal SD of 10 mmHg. The observed treatment effect, d, is therefore: $x-y=(X+e_X)-(Y+e_Y)=D+e_x-e_y$. Figure 2 shows a scatter plot for x and d with a sample size of 5000 generated within R. When d is regressed on x, the regression coefficient is 0.302, which is biased away from zero due to e_X. Oldham's method shows that the correlation between $x-y$ and $(x+y)/2$ is -0.002 and the regression coefficient, where $x-y$ is the dependent variable and $(x+y)/2$ the independent variable, is -0.001, indicating there to be no genuine differential treatment effects. Blomqvist's formula [17], given in equation (1), shows that the corrected regression

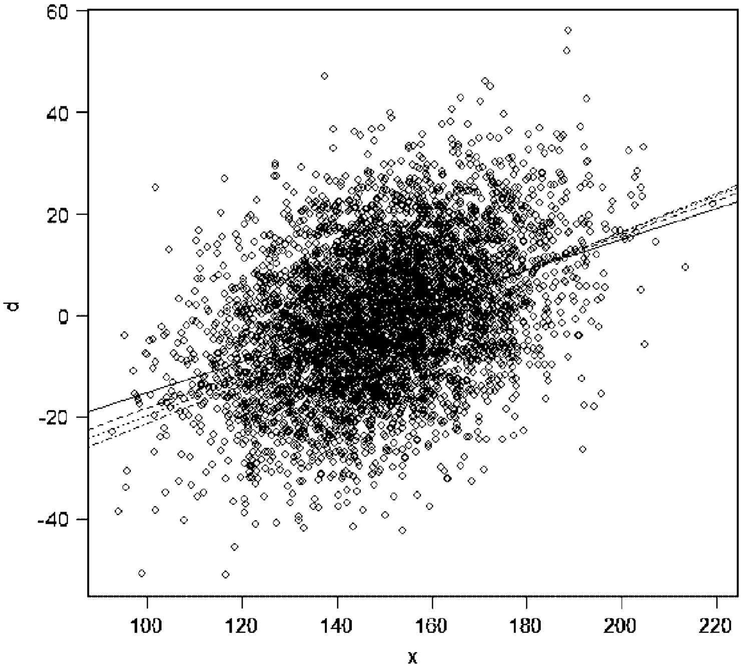

Figure 2. Scatter plot of observed systolic blood pressure reduction and observed baseline systolic blood pressure; sample size $=5000$. The overall regression slope (solid line) is 0.302. For the subgroup with greater than 160 mmHg baseline systolic blood pressure (sample size $=1372$) the regression slope is 0.341 (dashed line). For the subgroup with greater than 165 mmHg baseline systolic blood pressure (sample size $=971$) the regression slope is 0.360 (dotted line). For the group with greater than 170 mmHg of baseline systolic blood pressure (sample size $=640$) the regression slope is 0.376 (dot–dashed line).

slope is -0.009 (as $k=10^2/(15^2+10^2)=0.308$), which is very close to zero, indicating no genuine relationship between the true treatment effect (D) and initial SBP (X).

When the simulation is repeated 10 000 times, the median regression coefficient for x is 0.308 (2.5 and 97.5 centiles: 0.288, 0.327), and the corrected sloped is therefore zero. The median correlation coefficient between $x-y$ and $(x+y)/2$ is zero (2.5 and 97.5 centiles: -0.028, 0.028), and the median regression slope for $x-y$ regressed on $(x+y)/2$ is also zero (2.5 and 97.5 centiles: -0.023, 0.024). Thus, both methods give rise to correct analyses.

4.2. Scenario 1: individuals selected on the basis of high initial values

Now we reanalyse the data by selecting: (i) 1372 patients with initial blood pressure greater than 160 mmHg; (ii) 971 patients with initial blood pressure greater than 165 mmHg; or (iii) 640 patients with initial blood pressure greater than 170 mmHg. The regression coefficients for d regressed on x are 0.341 in (i), 0.360 in (ii), and 0.376 in (iii); Oldham's method yields negative correlations of -0.437 in (i), -0.494 in (ii), and -0.526 in (iii); and the regression slopes associated with Oldham's method are -0.576 in (i), -0.691 in (ii), and -0.762 in (iii), respectively. It should be noted that in Blomqvis's formula, k should in theory be derived from external estimation of error variance (see Section 3.1), which is not to be derived from the sample and is not based on selection of initial values. Hence the value of $k = 0.308$ (calculated in Section 4.1) shall be applied throughout this study in all scenarios. Blomqvist's formula then shows that the corrected regression slopes are 0.048 in (i), 0.075 in (ii), and 0.098 in (iii). Since there is genuinely no underlying relationship between baseline and change, Oldham's and Blomqvist's approaches are expected to give zero correlations or zero regression slopes. The departure from zero in Blomqvist's formula seems to be much smaller than in Oldham's method.

Results of 10 000 simulations show that the median regression slopes are 0.307 (2.5 and 97.5 centiles: 0.234, 0.380) in (i), 0.307 (2.5 and 97.5 centiles: 0.213, 0.402) in (ii), and 0.306 (2.5 and 97.5 centiles: 0.179, 0.433) in (iii); and the corrected regression coefficients using Blomqvist's formula are −0.001 in (i), −0.001 in (ii), and −0.003 in (iii). It should be noted that the reason that the medians of the corrected regression coefficients are all very close to zero is because the assumption of normally distributed x and y values is upheld by the simulation process, and under such 'ideal' circumstances the corrected slope for the truncated data will always be close to that for the full data set. However, slight departures from the assumption of normality could lead to a very different median corrected value. In other words, whilst Blomqvist's formula performs well under ideal circumstances, the correction is more sensitive to the normality assumptions for the truncated data than for the full data set. Furthermore, as the range of 2.5 and 97.5 centiles of the corrected slopes increase by selecting subgroups of higher baseline values, this represents an increase in the random variation due to smaller samples within subgroups. In contrast, Oldham's method always gives rise to a biased negative correlation, and in the simulations the correlation coefficients were −0.464 (2.5 and 97.5 centiles: −0.507, −0.418) in (i), −0.515 (2.5 and 97.5 centiles: −0.564, −0.462) in (ii), and −0.562 (2.5 and 97.5 centiles: −0.620, −0.499) in (iii); with corresponding regression slopes of −0.609 (2.5 and 97.5 centiles: −0.679, −0.539) in (i), −0.706 (2.5 and 97.5 centiles: −0.794, −0.619) in (ii), and −0.802 (2.5 and 97.5 centiles: −0.912, −0.690) in (iii).

By selecting a sub-group with greater baseline values than the whole sample, variations in the observed biases in the association between d and x, due to measurement error, became greater (Figure 2) and Blomqvist's method will show a positive or negative difference from zero due to under- or over-adjustment. Nevertheless, by selecting a sub-group with greater baseline values, the variance of the sub-group post-treatment values will be greater than that of the sub-group baseline values; Oldham's method therefore shows a spurious inverse association between treatment effect and baseline. In this scenario, Blomqvist's formula seems to perform better than Oldham's method.

4.3. *Scenario 2: true treatment effects differ across individuals*

Suppose now that the true treatment effect varies among patients with the same true baseline value. Following our previous notation, D is now a random variable with mean 20 mmHg and SD of 10 mmHg, and under the assumption that the expected true treatment effects are not related to baseline value, the correlation between X and D is expected to be zero. Regression analysis shows the estimated slope for change $(x - y)$ regressed on baseline (x) is 0.290. The corrected regression slope using Blomqvist's formula is −0.026, suggesting that the extent of blood pressure reduction is not related to unobserved baseline SBP. Given any true baseline SBP, the expected change in blood pressure is close to zero because the probability of obtaining a positive change or a negative change is almost equal. This is very different from the scenario where the expected change in blood pressure is constant. In contrast, Oldham's method yields a moderate negative correlation ($r = -0.177$ and regression coefficient $= -0.174$) between the observed change and average. The reason for a negative correlation obtained using Oldham's method is that the variance of y is greater than that of x because, whilst the variance of x is equal to the sum of two variances (namely, that of X and e_X), the variance of y is the sum of three variances (namely, that of X, e_Y and D, since $y = X - D + e_Y$). If the variances of e_X and e_Y are similar, the variance of y will be greater than that of x. The results of 10 000 simulations show that the median of regression slope is 0.308 (2.5 and 97.5 centiles: 0.238, 0.332), and the corrected regression slope is therefore zero. Both the median correlation coefficient and regression slope are -0.167 using Oldham's method.

Hayes [15] and others [11] have argued (incorrectly) that when true treatment effects differ across subjects, Oldham's method would give rise to a misleading association when the true treatment effect is *not* associated with true baseline values. The expected zero correlation or regression coefficient between the unobserved true baseline values X and true treatment effects D, as given by Blomqvist's approach, is *misinterpreted* as evidence to show that there is no *differential* treatment effect across the levels of baseline values. As Oldham pointed out in correspondence with his critics [28, 29], the relationship between $D = X - Y$ and X is potentially deceiving and should be interpreted cautiously. In our simulations, the zero correlation between D and X does not prove that there are no *differential* treatment effects, i.e. for greater baseline values, greater treatment effects will be achieved. On the contrary, there is a reverse baseline effect (i.e. for greater baseline values, lesser treatment effects will be achieved), because the variance of y is greater than that of x.

As first pointed out by Oldham [29], the correct interpretation of the zero correlation between D and X is that, due to the response of the patients to treatment being so *heterogeneous*, the correlation between D and X becomes zero. Moreover, the variation in treatment effects is the same across all levels of baseline value (i.e. given any baseline value of X, the expected treatment effect, D, always has a mean of 20 mmHg and SD of 10 mm); which is the reason why the correlation between D and X is close to zero. However, it is the difference in variances between the post-treatment value Y and the baseline value X that is crucial to the interpretation of the relationship between treatment effects and baseline values, not the correlation or regression slope between treatment effect and baseline. The correct interpretation of Scenario 2 is that baseline blood pressure is a poor predictor for blood pressure reduction after the treatment, but, the increased variance of post-treatment values indicates that, in general, individuals with lower baseline blood pressure respond better to the treatment than those with higher blood pressure.

Another way to reveal that the expected zero regression coefficient given by Blomqvist's formula in scenario (2) should not be interpreted as evidence of *no* differential baseline effect on the treatment is to regress change in blood pressure on the post-treatment blood pressure. Since the true treatment effect D is unrelated to the true baseline blood pressure X, there is also no relation between D and the true post-treatment blood pressure Y. However, the corrected regression slope (uncorrected slope $= -0.464$) given by Blomqvist's formula for d regressed on y (opposed to x) is -0.225 (opposed to -0.026). Results from 10 000 simulations show that the median regression slope for d regressed on y is -0.471 (2.5 and 97.5 centiles: -0.490, -0.451), and the corrected slope by Blomqvist's formula is -0.236. *Any statistical* *method to test differential treatment effects should yield equivalent results testing the relation between change and either initial value or final value*; *the corrected regression slope should be equivalent but in opposite directions.*

5. OLDHAM'S METHOD AND BLOMQVIST'S FORMULA ANSWER TWO DIFFERENT QUESTIONS

The misinterpretation of different results between Oldham's method and Blomqvist's formula in Scenario 2 is due to overlooking the full impact of regression to the mean in testing the relation between change and initial value. In the models proposed by Healy [24] and Hayes [15], and the explanation given by others [7, 8, 11], only measurement error (e_X) in the true initial value X is considered to be the cause of regression to the mean. Therefore, any method to correct for the bias caused by e_X, such as Blomqvist's formula, is (mistakenly) believed to be a sufficient solution. Blomqvist's formula provides an unbiased estimate, within regression analysis, of the treatment effect (e.g. change in blood pressure) *conditional* on an initial value (e.g. initial blood pressure). The regression slope is biased due to measurement error and/or biological variation in the measurement of initial values (i.e. e_X). By estimating the magnitude of measurement error in baseline values, Blomqvist's formula gives the correct estimate of regression slope. However, another cause of regression to the mean is the heterogeneous response to treatment, which does not bias the estimate of regression slope, and therefore will not be corrected by Blomqvist's formula. Blomqvist's formula is valid only in estimating the unbiased regression slope; it does not answer the question of whether or not there is a differential baseline effect, addressed by Oldham's method (and other approaches as discussed).

In summary, Blomqvist's formula gives an unbiased estimate of how much change is achieved *given* a baseline value. However, a non-zero regression slope *cannot* be interpreted as meaning there is a differential baseline effect. Therefore, Blomqvist's formula aims to correct the biased regression *slope* due to error in the initial values. Oldham's method, on the other hand, gives an unbiased test of *correlation* for differential baseline effects, as testing the correlation between $x-y$ and x is potentially misleading; though Oldham's method cannot yield any inference on change conditional on initial value. To help clarity this crucial distinction, it is useful to briefly revisit how Galton first discovered regression to the mean more than one century ago.

6. WHAT IS GALTON'S REGRESSION TO THE MEAN?

Galton wanted to study the heritance of human intelligence. However, owing to the lack of a precise measure of intelligence, he turned to measurable traits, such as body height [44, 45]. He invited families to his laboratory to measure their body heights. As males on average are taller than females, all female heights were multiplied by 1.08, and then he plotted the average of both parents' heights against their off-spring's heights [46, 47]. To his surprise, he found that, although adult children of tall parents were still taller than most people, they were generally shorter than their parents, i.e. they were closer to the mean height of the population. On the other hand, adult children of short parents, whilst still short, were on average taller than their parents, i.e. they were closer to the mean height of the population. Galton named this phenomenon *regression toward mediocrity*, and we know it today as regression to the mean.

It is important to note that regression to the mean in Galton's study was not the relation between repeated measurements of heights on the *same* individuals (where regression to the mean might occur, though is generally quite small), but the relation between measurements of body heights *across generations* (i.e. between parents and their adult children). Consequently, in Galton's study, regression to the mean was not only caused by measurement errors of individuals' heights, but also caused by the underlying genetic and environmental factors related to the growth of body height. Suppose that all parents' and their adult children's heights were measured twice, one week apart. This would provide information about the magnitude of measurement error and/or biological fluctuation in height (though the latter is probably ignorable). However, this would not eliminate regression to the mean in the analysis of the relation between body heights across generations. Compared to the variation in body heights occurring *across generations*, due to biological and/or environmental factors, the variation in measurement errors would be small and have modest impact on the correlation between heights of parents and their adult children.

In Galton's study, regression to the mean did not only occur at the individual level, but also at the population level (i.e. across generations). Just like our previous scenarios on the relationship between change and initial value following treatment, regression to the mean does not only occur due to measurement error in the initial value, but also due to the heterogeneous response of individuals to intervention. In statistical language, the non-unity correlation between the observed initial value x and observed post-treatment y is caused by both the measurement errors (and/or biological variation) in x and y, *and* individuals' heterogeneous responses to the treatment; both are regression to the mean.

7. CONCLUDING REMARKS

The misunderstanding of differences between Blomqvist's and Oldham's approaches not only impacts upon the testing of the relation between change and initial value. For instance, consider the debate regarding how to investigate the underlying risk as a source of heterogeneity in meta-analyses [48–52]. The approach of Oldham's method has been criticized for yielding misleading results when there is variation in the treatment effects across different clinical trials [50–52]. However, this criticism seems to be based on the same arguments used by Hayes to criticize Oldham's method in testing the relation between change and initial value in the Scenario 2; and we have shown this criticism to be untenable. If the treatment really works better in the trials with greater underlying risk of developing diseases, the variance of log odds in the treatment groups will be smaller than the variance of log odds in the placebo groups. If these two variances remain similar, the treatment effects are not related to the underlying risk in the placebo groups. Certainly, the problem with meta-analyses is more complex [51], since sample sizes and protocols are usually different across different trials.

Another conceptual confusion is around the distinction between mathematical coupling and regression to the mean [39]. Mathematical coupling is more often encountered in clinical research. For instance, within anaesthesiology and critical care on the oxygen consumption and oxygen delivery [53], both indices are derived using complex formula with some common components. When the shared components in both indices are measured with error, testing their inter-relationship using correlation or regression yields biased results, just as measurement errors in baseline values causes biased estimates in the testing of a relation between change and initial value. However, when both are measured without error, mathematical coupling still prevails. Whilst approaches to correct for the errors can give rise to an unbiased estimate of regression slopes [7, 8], for coupled variables, such methods cannot estimate the 'strength' of the relation (i.e. whether or not a relation is statistically significant), because the null hypothesis might no longer be zero. This more general situation is analogous to using Blomqvist's formula to correct for the bias in the relation between change and initial value. Correcting for measurement errors in coupled variables might yield unbiased estimates of oxygen consumption for a given level of oxygen delivery, but it remains uncertain whether or not the relation between oxygen consumption and delivery is statistically significant.

In the relation between change and baseline, mathematical coupling and regression to the mean are almost synonymous, as x and y are two repeated measurements. In a more general scenario, such as for oxygen delivery and oxygen consumption, the problem of mathematical coupling between variables is not limited to measurement error, but also the testing of an inappropriate null hypothesis. This is perhaps why, when considering only repeated measures as an illustration, there has been a poor distinction between mathematical coupling and regression to the mean. Consequently, Oldham's method has been misunderstood for many years, and Blomqvist's formula has been recommended, incorrectly, as a solution to a problem that it is not able to answer. Both Oldham's method and Blomqvist's formula are valid when they are applied to appropriate research questions, and both have their limitations.

ACKNOWLEDGEMENTS

The authors would like to thank Dr Samuel Manda for his comments on an earlier version of this manuscript and the three reviewers for their very helpful suggestions.

REFERENCES

1. Gill JS, Zezulka AV, Beevers DG, Davies P. Relation between initial blood pressure and its fall with treatment. *Lancet* 1985; **1**:567–568.
2. Halverson JD, Koehler RE. Gastric bypass: analysis of weight loss and factors determining success. *Surgery* 1981; **90**:446–455.
3. Blomqvist N. On the relation between change and initial value. *Journal of the American Statistical Association* 1977; **72**:746–749.
4. Jin P. Toward a reconceptualization of the law of initial value. *Psychological Bulletin* 1992; **111**:176–184.
5. Archie JP. Mathematical coupling: a common source of error. *Annals of Surgery* 1981; **193**:296–303.
6. Andersen B. *Methodological Errors in Medical Research*. Blackwell: London, 1990.
7. Moreno LF, Stratton HH, Newell JC, Feustel PJ. Mathematical coupling of data: correction of a common error for linear calculations. *Journal of Applied Physiology* 1986; **60**:335–343.
8. Stratton HH, Feustel PJ, Newell JC. Regression of calculated variables in the presence of shared measurement error. *Journal of Applied Physiology* 1987; **62**:2083–2093.
9. Tu Y-K, Gilthorpe MS, Griffiths GS. Is reduction of pocket probing depth correlated with the baseline value or is it 'mathematical coupling'? *Journal of Dental Research* 2002; **81**:722–726.
10. Tu Y-K, Maddick IH, Griffiths GS, Gilthorpe MS. Mathematical coupling still undermines the statistical assessment of clinical research: illustration from the treatment of guided tissue regeneration. *Journal of Dentistry* 2004; **32**:133–142.
11. Kirkwood B, Sterne JAC. *Medical Statistics* (2nd edn). Blackwell: Oxford, 2003.
12. Oldham PD. A note on the analysis of repeated measurements of the same subjects. *Journal of Chronic Diseases* 1962; **15**:969–977.
13. Altman DG. Statistics in medical journals. *Statistics in Medicine* 1982; **1**:59–71.

14. Altman DG. Statistics in medical journals: developments in the 1980s. *Statistics in Medicine* 1991; **10**:1897–1913.
15. Hayes RJ. Methods for assessing whether change depends on initial value. *Statistics in Medicine* 1988; **7**:915–927.
16. Altman DG. *Practical Statistics for Medical Research*. Chapman & Hall/CRC: London, Boca Raton, FL, 1991.
17. Blomqvist N. On the bias caused by regression towards the mean. *Journal of Clinical Periodontology* 1987; **14**:34–37.
18. Yudkin PL, Stratton HH. How to deal with regression to the mean in intervention studies. *Lancet* 1996; **347**:241–243.
19. Fitzmaurice G. Regression to the mean. *Nutrition* 2000; **16**:81–82.
20. Morton V, Torgerson DJ. Effect of regression to the mean on decision making in health care. *British Medical Journal* 2003; **326**:1083–1084.
21. Bland JM, Altman DG. Regression towards the mean. *British Medical Journal* 1994; **308**:1499.
22. Bland JM, Altman DG. Some examples of regression towards the mean. *British Medical Journal* 1994; **309**:780.
23. Campbell DT, Kenney DA. *A Primer on Regression Artifacts*. Guilford Press: New York, 2003.
24. Healy MJR. Some problems in repeated measurements. In *Perspectives in Medical Statistics*, Bithell JF, Coppi R (eds). Academic Press: London, 1981.
25. Pitman E. A note on normal correlation. *Biometrika* 1939; **31**:9–12.
26. Morgan W. A test for the significance of the differences between two variances in a sample from a normal bivariate distribution. *Biometrika* 1939; **31**:13–19.
27. Bland JM, Altman DG. Statistical methods for assessing agreement between two methods of clinical measurement. *Lancet* 1986; **327**:307–310.
28. Garside RF. Letter to the editor: on the analysis of repeated measurements of the same subjects. *Journal of Chronic Diseases* 1963; **16**:445–449.
29. Oldham PD. Letter to the editor. *Journal of Chronic Diseases* 1963; **16**:447–450.
30. Fisher RA. Frequency distribution of the values of the correlation coefficient in samples from an indefinitely large population. *Biometrika* 1915; **10**:507–521.
31. Greenen R, van de Vijver FJR. A simple test of the law of initial value. *Psychophysiology* 1993; **30**:525–530.
32. Guilford JP, Fruchter B. *Fundamental Statistics in Psychology and Education*. McGraw-Hill: New York, 1973.
33. Maloney CJ, Rastogi SC. Significance test for Grubb's estimators. *Biometrics* 1970; **26**:671–676.
34. Myrtek M, Foerster F. The law of initial value: a rare exception. *Biological Psychology* 1986; **22**:227–237.
35. Dunn G. *Statistical Evaluation of Measurement Errors* (2nd edn). Arnold: London, 2004.
36. Cleary PJ. L.I.V. R.I.P.?: comments on Myrtek and Foerster's 'the law of initial value: a rare exception'. *Biological Psychology* 1986; **22**:279–284.
37. Selvin S. *Practical Biostatistical Methods*. Duxbury: Pacific Grove, 1994.
38. Andersen B. *Methodological Errors in Medical Research*. Blackwell: London, 1990.
39. Bartko JJ, Pettigrew K. A note on the correlation of parts with wholes. *American Statistician* 1968; **22**:41.
40. Tu Y-K, Baelum V, Gilthorpe MS. The relationship between baseline value and its change: problems in categorization and the proposal of a new method. *European Journal of Oral Sciences* 2005; **113**:279–288.
41. Dawson B, Trapp RG. *Basic and Clinical Biostatistics* (3rd edn). McGraw-Hill: New York, 2001.
42. Goldstein H. *Multilevel Statistical Models* (2nd edn). Wiley: New York, 1995.
43. Blance A, Tu Y-K, Gilthorpe MS. A multilevel modelling solution to mathematical coupling. *Statistical Methods in Medical Research* 2005; **14**:553–565.
44. Galton F. Regression toward mediocrity in hereditary stature. *Journal of Anthropological Institute of Great Britain and Ireland* 1986; **15**:246–263.
45. Stigler SM. Regression towards the mean, historically considered. *Statistical Methods in Medical Research* 1997; **6**:103–114.
46. Senn SJ. *Dicing with Death*. Cambridge University Press: Cambridge, MA, 2003.
47. Hanley JA. 'Transmuting' women into men: Galton's family data on human stature. *American Statistician* 2004; **58**:237–243.
48. Senn SJ. Letter to the editor: importance of trends in the interpretation of an overall odds ratio in the meta-analysis of clinical trials. *Statistics in Medicine* 1994; **13**:293–295.
49. Brand R. Author's reply: importance of trends in the interpretation of an overall odds ratio in the meta-analysis of clinical trials. *Statistics in Medicine* 1994; **13**:295–296.
50. Sharp SJ, Thompson SG, Altman DG. The relation between treatment benefit and underlying risk in meta-analysis. *British Medical Journal* 1996; **313**:735–738.
51. Van Houwelingen H, Senn S. Letter to the editor. *Statistics in Medicine* 1999; **18**:110–113.
52. Thompson SG, Smith TC, Sharp SJ. Investigating underlying risk as a source of heterogeneity in meta-analysis. *Statistics in Medicine* 1997; **16**:2741–2758.
53. Myles PS, Gin T. *Statistical Methods for Anaethesia and Intensive Care*. Butterworth and Heinemann: Oxford, 2000.