A solution to dependency: using multilevel analysis to accommodate nested data

Supplemental simulation and analysis

Emmeke Aarts, Matthijs Verhage, Jesse V. Veenvliet, Conor V. Dolan, Sophie van der Sluis

Supplemental simulation

Simulations and analyses were conducted in R64 2.11.111 (R Development Core Team. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria (2010)). In all simulations, a standardized (i.e., all variables have a mean of 0 and standard deviation of 1), two-level multilevel model was used, where individual observations i are nested within clusters j. To clarify the simulations, we use an example in which the individual observations are made on cells, and the clusters in which the cells are nested are mice. In our example, we wish to test whether cells from wild type (WT) and knockout (KO) mice differ with respect to a continuous and normally distributed characteristic of the cell, Y. Note that WT vs. KO (i.e., genotype) is our dichotomous experimental variable, X, and X only varies over, not within, mice (i.e., cells harvested from one mouse always have the same genotype). The multilevel model for the present example is given by:

$$Y_{ij} = \beta_{0j} + \beta_1 X_j + e_{ij}, \tag{1}$$

i.e., the outcome variable Y for cell i from mouse j (Y_{ij}) is regressed on the experimental variable value of mouse j (X_j), with the mouse specific intercept β_{0j}, slope parameter β_1, and the zero mean residual (prediction error) e_{ij}. The variance of e, i.e., σ_e^2, is referred to as the residual variance. The mouse specific intercept is given by:

$$\beta_{0j} = \gamma_{00} + u_{0j}, \tag{2}$$

where γ_{00} is the overall intercept (i.e., mean) across all mice, and u_{0j} is the mouse-specific deviation from the overall intercept. The variance of u_{0j}, i.e., $\sigma_{u_0}^2$, is referred to as intercept variance and equals the ICC in the standardized model. As we are interested in the difference in Y between genotypes, the null hypothesis H-null: $\beta_1 = 0$ (Note that if we were to pool all observations, and discard the mouse-specific information, testing H-null is: $\beta_1 = 0$ would correspond to the conventional Student t-test). See Figures 2 and 3 in the main text for a graphical illustration of the ICC and a graphical representation of Students t-test and the multilevel model, respectively.
The random datasets in the simulations were generated as follows. For each parameter setting, we generated 10.000 simulated datasets. In all data sets, the overall intercept γ_{00} was set to 4 (i.e., an arbitrary value as the overall intercept does not influence the test of β_1). As a standardized model with equal group sizes was used, the predictor variable X was coded -1 and 1. Since X is coded -1 and 1 (and not the usual dummy coding of 0 and 1), the regression slope β_1 equals half the difference between experimental groups given by the effect size Cohen's d (Cohen, J. Statistical power analysis for the behavioral sciences (Erlbaum, Hillsdale, JN, 1998); the effect size Cohen's d was set to either 0.20, 0.50 or 0.80, see below).

Nature Neuroscience: doi:10.1038/nn.3648

The cluster specific zero mean deviations from the general intercept μ_{0j} were obtained from a multivariate normal distribution with a mean of zero and variance equal to the unexplained ICC. The Cohen's d was transformed to the metric of explained variance to obtain the explained ICC by:

$$R^2 = \left(\frac{d}{\sqrt{d^2 + 4}} \right)^2. \quad (3)$$

As we used a standardized model, σ_e^2 was set such that σ_e^2 + unexplained ICC + explained ICC = 1 in the intercept-only model. The residual variance σ_e^2 was generated separately from a normal distribution with a mean of zero and variance of σ_e^2 ($\sim N(0, \sigma_e^2)$).

Inflated Type I error rate (Fig. 1a)

To illustrate how the Type I error rate changes as function of the number of observations per cluster and the magnitude of the (unexplained) ICC, we simulated data in which the experimental effect β_1 equaled zero. The number of observations per cluster ranged from 5 to 105 with an increment of 2. The ICC was set at 0.10 or 0.50. Per parameter setting, 10,000 data sets were generated that were analyzed using either a conventional t-test, or a multilevel model. For each type of analysis, we counted the number of times out of 10,000 that the regression slope (β_1) was statistically significant given $\alpha = 0.05$.

Loss in power (Fig. 1b)

To illustrate the loss in power when using conventional t-tests on cluster based summary statistics compared to multilevel analysis on all observations, we simulated data in which the parameter β_1 equaled either a small, medium or large experimental effect according to Cohen's d ($d = 0.20$, 0.50 or 0.80, respectively). The number of clusters ranged from 10 to 80 with an increment of 2. The unexplained ICC was set at 0.10 or 0.50. For each parameter setting, the 10,000 data sets were analyzed using either multilevel analysis on all individual observations, or conventional analysis on the cluster means (i.e., the mean value per cluster was calculated followed by an independent samples t-test to obtain the statistical significance of the experimental effect). For both types of analyses, we then counted the number of times that the experimental effect β_1 was statistically significant given $\alpha = 0.05$. The loss in power was calculated by subtracting the relative power of the conventional t-test from 1, where the relative power is obtained by dividing the estimated power associated with the t-test on summary statistics by the estimated power of multilevel analysis on individual observations.

Nature Neuroscience: doi:10.1038/nn.3648

Supplemental analysis

To clarify the procedure of a multilevel analysis, we will use a hypothetical example in which measurements of dendrite length (*Length*) are nested within neurons (*Neuron_ID*). In this example, we investigate whether dendrite length differs between knock out and wild type mice (*Genotype*). Collected from 15 cells (7 wild type and 8 knockout, respectively), there are in total 292 observations of the outcome variable *Length*. Note that this is not a large sample for multilevel-analysis. However, our model is relatively simple (just one predictor, *Genotype*, and one outcome variable, *Length*), so we do not expect problems with model convergence. Convergence problems could, however, occur when running more complex models (e.g., multiple predictors, inclusion of covariates) on a dataset of this size.
Before we run the analysis, we standardized all variables (i.e., all variables have a mean of 0 and a standard deviation of 1). This is convenient because in a standardized model the between group variance approximates the ICC, and the slope for *Genotype* represents Cohens d divided by 2. Standardized variables can easily be obtained in SPSS (*Analyze* $\Rightarrow$ *Descriptive Statistics* $\Rightarrow$ *Descriptives*: select the variables you want to standardize and tick the box Save standardized values as variables). When the number of observations in each experimental condition is equal, wild type and knockout are coded as -1 and 1, respectively, to standardize *Genotype*. In this example, group sizes are not equal so the standardized values for *Genotype* are -.99482 and 1.00173, respectively.

Assumptions

One of the assumptions of standard multilevel analysis is that the outcome variable is normally distributed. A visual inspection of the distribution of *Length* shows that *Length* can be considered normally distributed. When data are non-normal, transformations can be considered, or a model for non-normal data can be used (see below). When the results of multilevel analysis of the transformed and untransformed data are similar, interpreting the results of the untransformed data can be easier, and is therefore recommended.
Another assumption concerns the absence of outliers, i.e., standardized values below -3 and above 3 need to be excluded from the analysis.

Analysis

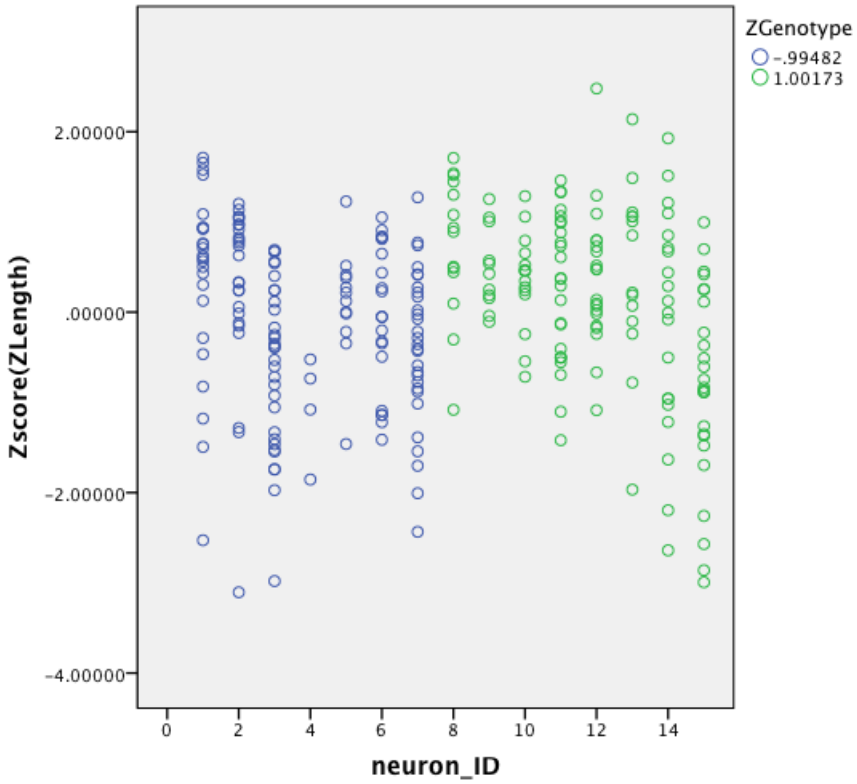
Before conducting the actual multilevel analysis, we will first visually examine the variance over and within neurons of the values of the dendrite length to get an idea of the degree of relative similarity between observations obtained from the same neuron. Also, visual inspection will allow a first assessment of whether the variation between cells is caused by random variation only, or (partly) by genotype as well. We plot the values of *Length* for each *Neuron_ID* separately and color code the distinct observations from WT and KO: syntax and output are shown in Supplementary Table 2. The figure shows that there is quite some variation both within and between neurons, which cannot all be explained by *Genotype*. In general, however, the mean dendrite length seems slightly higher for KO mice, but we of course need to test this.

In order to perform the analysis, one additional variable has to be created: an artificial intercept (*int*), which is a variable that always has value 1. Next, an estimate of the intracluster correlation (ICC) can be obtained by running an intercept only model, i.e., a model in which every neuron is allowed to have its own mean dendrite length (see Figure 3b in the manuscript), but that does not include *Genotype* as predictor: syntax and selected output are presented in Supplementary Table 3.

Nature Neuroscience: doi:10.1038/nn.3648

Supplementary Table 2. Syntax and selected output for visualization of variance of *Length*

```
GRAPH
/SCATTERPLOT(BIVAR)=neuron_ID WITH ZLength BY ZGenotype
/MISSING=LISTWISE.
```



To obtain an estimate of the ICC, apply the equation given in Figure 2a:

$$ICC = \frac{\sigma_u^2}{\sigma_u^2 + \sigma_e^2} = \frac{.1685}{.1685 + .8284} = .169. \quad (4)$$

This means that 16.9% of the variability in *Length* is due to differences between neurons, i.e., can be explained by neuron-membership. As the simulations presented in Figure 1a showed that the Type I error rate can already be much increased when ICC = .10, multi-level analysis is certainly advised when ICC = .169. Note that because we use standardized variables only, the variance of the intercept (which can be interpreted as the between cluster variance) in the table approximates the value of the ICC: the slight deviation in the third decimal is due to the fact that *Length* is not perfectly normally distributed.

If we want to test whether the ICC is significantly different from 0, we thus test in a standardized situation whether the variance of the intercept is significantly different from 0. However, the Wald test reported in the table is not appropriate to test significance of variances (i.e., the asymptotic Wald test assumes normally distributed variance components, which is unrealistic; Bryk, A.S. & Raudenbush, S.W. Hierarchical Linear Models (Sage, Newbury Park, CA, 1992)). Whether the

Nature Neuroscience: doi:10.1038/nn.3648

Supplementary Table 3. Syntax and selected output for Intercepts-only model through SPSS MIXED

```
MIXED
ZLength with int
/fixed int — noint
/random int — subject(Neuron_ID) covtype(un)
/METHOD = ML
/print solution testcov r.
```

Estimates of Fixed Effects[a]

Parameter	Estimate	Std. Error	df	t	Sig.	95% Confidence Interval	
						Lower Bound	Upper Bound
int	.038590	.120811	14.061	.319	.754	-.220418	.297598

a. Dependent Variable: ZLength Zscore(ZLength).

Covariance Parameters

Estimates of Covariance Parameters[a]

Parameter		Estimate	Std. Error	Wald Z	Sig.	95% Confidence Interval	
						Lower Bound	Upper Bound
Residual		.828392	.070835	11.695	.000	.700568	.979538
int [subject = neuron_ID]	Variance	.168516	.081332	2.072	.038	.065437	.433972

a. Dependent Variable: ZLength Zscore(ZLength).

variance component is significantly different from 0 can, however, be tested using a chi-square (χ^2) test. If we square the Wald Z statistic in the table, we get approximately a chi-square value, with the number of degrees of freedom being 1 (i.e., we test only 1 parameter, namely the variance of the intercept). So we get:

$$\chi^2 = (2.072)^2 = 4.293 \tag{5}$$

Note that since a variance component cannot be negative and this parameter is thus subject to boundary constraints (Berkhof, J. & Snijders, T. A. B. Variance Component Testing in Multilevel Models. J Educ Behav Stat 26, 133-152 (2001); Stoel, R.D., Garre, F. G. , Dolan, C. & van den Wittenboer, G. On the Likelihood Ratio Test in structural equation modeling when parameters are subject to boundary constraints. Psychol Methods 11, 439-455 (2006)), the accompanying p-value, which equals .038, needs to be divided by 2: $p = .019$. Assuming $\alpha = .05$, this test is significant, i.e., the variance of the intercept, and thus the ICC, is significantly different from 0 (note that SPSS prints -2 Log Likelihood information in the table with information criteria. Usually, this -2LL information is used to calculate the chi-square test. However, SPSS sometimes uses pseudo maximum likelihood estimation, and then the -2LL values of different models cannot be used to compute a chi-square value).

Now we have estimated the overall ICC, and have determined that the ICC is significant, we will add *Genotype* to our model to identify the effect of *Genotype*, i.e., is the dendrite length different between wild type and knockout mice, to determine how much between-neuron variation

Nature Neuroscience: doi:10.1038/nn.3648

is caused by *Genotype*, and to determine how much between-neuron variation cannot be explained by *Genotype* but is thus due to other, unknown, causes. The syntax and selected output is presented in Supplementary Table 4.

Supplementary Table 4. Syntax and selected output for model including the predictor *Genotype* through SPSS MIXED

```
MIXED
ZLength with int ZGenotype
/fixed int ZGenotype — noint
/random int — subject(Neuron_ID) covtype(un)
/METHOD = ML
/print solution testcov r.
```

Estimates of Fixed Effects[a]

Parameter	Estimate	Std. Error	df	t	Sig.	95% Confidence Interval	
						Lower Bound	Upper Bound
int	.025935	.115299	14.710	.225	.825	-.220240	.272111
ZGenotype	.166198	.115468	14.712	1.439	.171	-.080336	.412733

a. Dependent Variable: ZLength Zscore(ZLength).

Covariance Parameters

Estimates of Covariance Parameters[a]

Parameter		Estimate	Std. Error	Wald Z	Sig.	95% Confidence Interval	
						Lower Bound	Upper Bound
Residual		.826775	.070595	11.712	.000	.699370	.977389
int [subject = neuron_ID]	Variance	.148274	.071707	2.068	.039	.057466	.382575

a. Dependent Variable: ZLength Zscore(ZLength).

The standardized effect, i.e., Cohen's d, of *Genotype* can be calculated as $2 * .166 = .332$. The 95% confidence interval (CI) of the unstandardized effect of *Genotype* is obtained through $\beta_1 \pm Z_{1-\alpha} * SE_{\beta_1}$, which corresponds to $0.166 \pm 1.96 * 0.115 = [-0.060, 0.391]$. As the 95% CI contains the value 0, the effect of *Genotype* on the dendrite length is not significant, which is also indicated by the p-value of .171. Calculated as the between cell variance (i.e., intercept variance) of the model without *Genotype* (Supplementary Table 3) minus the between cell variance of the model with *Genotype* (Supplementary Table 4), *Genotype* explains only .1685 - .1483 = .020 of the between-cell variance.

The output of the analysis of the same data using a conventional t-test is shown in Supplementary Table 5. Although the estimate of the standardized effect is slightly lower than in the multilevel analysis ($2 * .125 = .250$), the effect of *Genotype* appears statistically significant (95% CI = [0.009, 0.241], $p = .033$) when the nesting of the data is not accommodated (i.e., when the data are treated as independent observations). The output of the t-test performed on the means calculated within each neuron, i.e., on the summary statistics, is shown in Supplementary Table 6. The effect of

Nature Neuroscience: doi:10.1038/nn.3648

Genotype is estimated slightly higher compared to the multilevel analysis ($2 * .204 = .408$), but with a 95% CI of [-0.011, 0.431] and a corresponding p-value of .138 it is not statistically significant. Note that in the dataset containing the summary statistics, *Genotype* has to be standardized again to have exactly a mean of 0 and a standard deviation of 1.

Supplementary Table 5. Selected output for conventional t-test on individual observations

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	1.032E-016	.058		.000	1.000
	ZGenotype	.125	.059	.125	2.141	.033

a. Dependent Variable: Zscore(ZLength)

Supplementary Table 6. Selected output for conventional t-test on summary statistics

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	.024	.128		.183	.857
	Zscore(ZGenotype)	.210	.133	.401	1.580	.138

a. Dependent Variable: ZLength_mean

Data and SPSS-syntax for this worked example can be downloaded from http://ctglab.nl/people/emmeke_aarts.

On this website, we also provide 4 additional worked examples of multilevel analysis online. These examples concern analysis of Type II data (i.e., data in which the predictor variable varies within objects, such that both the slope and the intercept can vary across objects), multilevel analysis of longitudinal data, and multilevel analysis of dichotomous and Poisson distributed outcome variables.

Nature Neuroscience: doi:10.1038/nn.3648