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

Emmeke Aarts[1], Matthijs Verhage[1,2], Jesse V. Veenvliet[3], Conor V. Dolan[4] and Sophie van der Sluis[1,5*]

[1*]Section Functional Genomics, Center for Neurogenomics and Cognitive Research, VU University Amsterdam, Amsterdam, The Netherlands.
[2]Section Functional Genomics, Department Clinical Genetics, VU Medical Center, Amsterdam, Amsterdam, The Netherlands.
[3]Center for Neuroscience, Swammerdam Institute for Life Sciences, Science Park, University of Amsterdam, Amsterdam, The Netherlands.
[4]Department of Biological Psychology, VU University Amsterdam, Amsterdam, The Netherlands.
[5]Section Complex Trait Genetics, Department of Clinical Genetics, VU Medical Center, Amsterdam, The Netherlands.

*Corresponding author(s). E-mail(s): s.vander.sluis@vu.nl;

Abstract

In neuroscience, experimental designs in which multiple observations are collected from a single research object (for example, multiple neurons from one animal) are common: 53% of 314 reviewed papers from five renowned journals included this type of data. These so-called 'nested designs' yield data that cannot be considered to be independent, and so violate the independency assumption of conventional statistical methods such as the t test. Ignoring this dependency results in a probability of incorrectly concluding that an effect is statistically significant that is far higher (up to 80%) than the nominal a level (usually set at 5%). We discuss the factors affecting the type I error rate and the statistical power in nested data, methods that accommodate dependency between observations and ways to determine the optimal study design when data are nested. Notably, optimization of experimental designs nearly always concerns collection of more truly independent observations, rather than more observations from one research object.

Neuroscience has seen major advances in understanding the nervous system over the past decades. Serious concerns have, however, been raised about an excess of false positive results contaminating the neuroscience literature[1–4]. Controlling the false positive rate is critical, since theoretical progress in the neuroscience field relies fundamentally on drawing correct conclusions from experimental research. Reported causes of increased levels of false positives range from inadequate sample size (i.e., underpowered studies), to lack of standardization with respect to research design, applied measures and corrections, exclusion/inclusion criteria, and choice of statistical methods. To improve transparency and reproducibility, Nature journals recently developed a checklist to aid authors to report basic methods information [5, 6]. Among things, authors are asked whether the assumptions of chosen statistical methods are met. Here, we show that one of these assumptions, i.e., the assumption of independent observations, is particularly relevant to neuroscience: neuroscience data often show dependency (i.e., "nesting", see Box 1 for definitions of key statistical terms) and failure to accommodate this is another, as yet neglected, cause of false positive results.

Nested designs are not unique to the neuroscience field but are also encountered, for instance, in the social sciences (e.g., children nested in classes, nested in schools), in behavioral genetics (e.g., relatives nested in families), and in the field of medicine (e.g., patients nested in doctors, nested in hospitals). In biomedical research, nested data are common in electron microscopy studies, with the n often at a subcellular level. In neuroscience, however, studies on neuron morphology and physiology typically give rise to nested data since technical advances allow researchers to obtain measurements on every dendrite of a neuron, and every spine of each dendrite, or to acquire multiple recordings of neuronal activity from the same cell.

1 The problem of nesting

Nested designs are designs in which multiple observations or measurements are collected in each research object (e.g., animal, tissue sample, or neuron/cell)[7]. Consider the following, fictive yet representative, research results: "The channel blocker significantly affected Ca^{2+} signals ($n = 120$ regions of interest (ROI) from 10 cells, $p < .01$)." "The number of vesicles docked at the active zone was smaller in presynaptic boutons in mutant neurons than in WT neurons ($n =$ 20 and 25 synapses each from 3 neurons for mutant and WT, $p < .01$)." Both statements concern experimental designs involving nested (or "clustered") data. These nested designs are particularly common to the neuroscience field, as many research questions in neuroscience consider multiple layers of complexity: from protein complexes, synapses, and neurons, to neuronal networks, connected systems in the brain, and behavior. In such multiple-layer-crossing designs, careful consideration of the issues that come with nesting is crucial to avoid incorrect inferences. The generality of nested designs in molecular, cellular, and developmental neuroscience is apparent from a literature study we

Box 1: Key statistical terms

Nested data Data that are characterized by a hierarchical or multilevel structure, i.e., organized at more than one level. In neuroscience, for instance, synapses (level 1) are organized, or nested, in cells (level 2).

Dependent observations Nesting often gives rise to dependency (i.e., similarity) among observations because observations obtained from the same research object (e.g., cell) tend to be more alike than observations taken from different objects. Most statistical tests assume observations to be independent. Violation of this independence-assumption can result in underestimated standard errors, underestimated p-values, and an increased Type I error rate.

Observed vs. effective sample size While independent observations convey unique information, dependent observations partly convey the same information. This loss of unique information reduces the observed sample size to the effective sample size, which denotes the number of independent observations required to carry the same amount of information as originally provided by the dependent ones.

Variance Estimate of the variability in a data set. In nested data, the total variance (VarT) is the sum of the variance within research objects (i.e., VarW: variability among observations taken from the same cell), and the variance between research objects (i.e., VarB: the variation in cell means).

Intracluster correlation (ICC) Index of the relative similarity of observations taken from the same research object (e.g., cell), and thus an indicator of the amount of dependence in the data. The ICC is calculated as VarB / [VarB + VarW]. Increasing the differences between research objects (VarB) and (or) decreasing the differences among measures within a research object (VarW), increases the ICC. Experimental manipulations (e.g., genotype) can increase the variability between objects (VarB) and thereby increase the ICC. The part of the ICC that can be attributed to the experimental manipulation is called the explained ICC. The remainder is called the unexplained ICC. In the this paper, we use the term ICC to indicate the unexplained ICC, unless stated otherwise.

Multilevel model A multilevel (aka nested, hierarchical linear, or random effects) model explicitly accommodates dependency between observations taken from the same object by allowing model parameters to differ between objects. By explicitly accommodating dependency, multilevel models consider the effective rather than the observed sample size, and as such prevent Type I error rate inflation.

Type I error or false positive The incorrect rejection of a true null-hypothesis. The probability to commit a Type I error is denoted by alpha (α), which is generally set at .05. Ignoring the nested structure of data may result in an inflated Type I error rate.

Type II error or false negative The failure to reject a false null-hypothesis. The probability to commit a Type II error is generally denoted by β.

Statistical power The probability to correctly reject a false null hypothesis, i.e., to detect an effect that is actually there. The power equals $1 - \beta$, where β denotes the probability to commit a Type II error.

Effect size An objective, standardized (i.e., scale free) measure of the magnitude of an observed effect. Cohen's d, for instance, represents the standardized difference between the means of two groups. In multilevel analysis, the explained ICC (i.e., the explained variance R^2) can be interpreted as effect size.

conducted involving research articles published over the last 18 months in *Science*, *Nature*, *Cell*, *Nature Neuroscience* and every first issue of the month of *Neuron* (see below): at least 53% of the 314 publications included nested data.

But why is nesting an issue? Since observations taken from the same research object (e.g., brain, animal, cell) tend to be more similar than observations taken from different objects (e.g., due to natural variation between objects, and differences in measurement procedures or conditions), nested designs yield clusters of observations that cannot be considered independent. Nevertheless, conventional statistical methods, like the *t*-test and ANOVA, are often used to analyze these nested data, even though these methods assume observations to be independent. Yet, the failure to take the dependency among observations into account forms a threat to the validity of the statistical inference. Depending on the number of observations per research object and the degree of dependence, the probability of incorrectly concluding that an effect is statistically significant (i.e., Type I error rate) can be far higher than the nominal level expressed by α (usually $\alpha = 0.05$). To illustrate the effect of nesting on results obtained through conventional tests, we conducted a simulation study (Fig. 1a, see Supplement 1.1 "Generation and analysis of simulated data" for details). Given a nominal α of 0.05, ignoring nesting can result in an actual

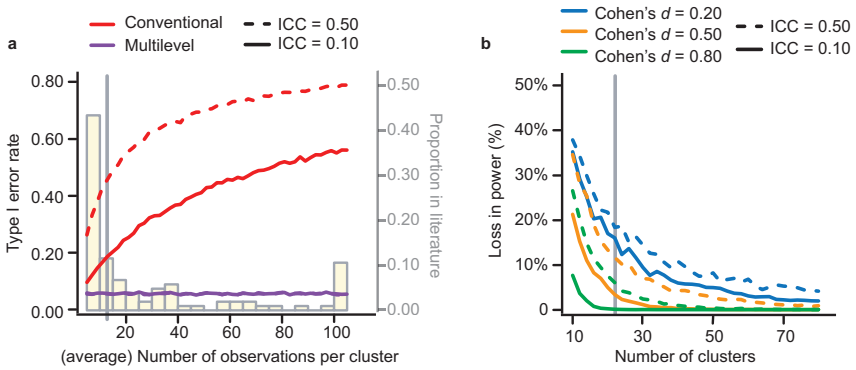

Fig. 1 Use of conventional t-test on nested data inflates the Type I error rate whereas cluster-based summary statistics decreases statistical power. In the background, the results of the literature study ($N = 314$) are shown. (**a**) Under two conditions (unexplained ICC = 0.10 or 0.50), nested data were simulated for two experimental groups (e.g., knockout versus wild type), with 25 clusters per group. The groups did not differ with respect to their means (i.e., no experimental effect). These nested data were analyzed using either a conventional t-test or multilevel analysis. When using a *t*-test, the Type I error increases steadily as the number of observations per cluster increases. In the background, the average number of observations per cluster is shown as observed in 314 research articles published in *Science*, *Nature*, *Cell*, *Nature Neuroscience*, and *Neuron*. The vertical grey line represents the median number of observations per cluster reported in the literature. (**b**) Under two conditions (unexplained ICC = 0.10 or 0.50), nested data were simulated for two experimental groups with a small, medium or large experimental effect (Cohen's *d* equals 0.20, 0.50 or 0.80, respectively). Compared to multilevel analysis, the loss in power when analyzing summary statistics is larger when the number of clusters is smaller. The vertical grey line represents the median number of clusters observed in the 7% published papers that reported analyses on cluster-based summary statistics where multilevel analysis could have been used.

Type I error rate as high as 0.80. That is, if no experimental effect is present, conventional methods that do not accommodate dependency will yield spurious statistically significant results in 80% of the studies (see Box 2 for a detailed discussion of the results and the theoretical proof).

To explain why clustering affects the Type I error rate, the distinction between the observed and the effective sample size is essential. The core of this distinction is: does each individual observation contribute unique information? This can be inferred from the degree of relative similarity between observations obtained from the same research object. This similarity is expressed in the intracluster correlation (ICC), which ranges from 0 to 1 (Fig. 2a–c). If clustering is absent (ICC = 0), all observations obtained from a research object are independent, i.e., contribute fully unique information. In the extreme case of ICC = 1, all observations obtained from the same research objects are equal and thus convey the very same information.

The experimental variable (e.g., genotype) can contribute to the dissimilarity of observations from different objects, and therefore to the relative similarity of observations from the same object. The part of the relative similarity that is attributable to the experimental variable is referred to as the explained ICC, while the part of the ICC that is attributable to other, unknown factors, is called the unexplained ICC. In the remainder of this paper, we use the term ICC to indicate the unexplained ICC, unless stated otherwise. Importantly, the unexplained part of the ICC causes inflation of the Type I error rate. In the extreme case that the ICC equals 1, the observed sample size may be N, but the effective sample size, i.e., the number of unique information units, equals the number of research objects (i.e., the number of clusters). For example, given 5 measurements on 10 cells, ICC = 0 implies a sample size of 5 $\times$10 = 50, but as the ICC tends to 1, the effective sample size tends to 10 (Fig. 2d). In terms of variation, correlation between observations from the same research objects (ICC > 0) reduces the variation in the total sample, compared to the variation expected in a random sample (ICC = 0; Fig. 2a). Because conventional statistical analyses are based on the observed rather than the effective sample size, standard errors of parameters are underestimated, and test statistics are overestimated. As a result, the associated p-values are too low, which results in excessive Type I error rate [8].

To correctly handle dependence in nested designs, multilevel models (also known as hierarchical or random effects models) can be used. These models produce correct Type I error rates (Fig. 1a). Alternatively, multilevel analysis can be circumvented by conducting conventional analyses on cluster-based summary statistics, e.g., by performing a t-test on the means/-medians calculated within each cluster. Although this strategy is statistically valid, information contributed by the individual observations is lost, and consequently, relative to multilevel analysis, statistical power to detect the experimental effect of interest decreases[7, 9, 10]. Conducting *t*-tests on cluster-based means instead of multilevel analysis on all observations, results in up to 40% loss of statistical power, depending on the number of clusters in the

study and the ICC (Fig. 1b; see Supplement 1.1. "Generation and analysis of simulated data" for details).

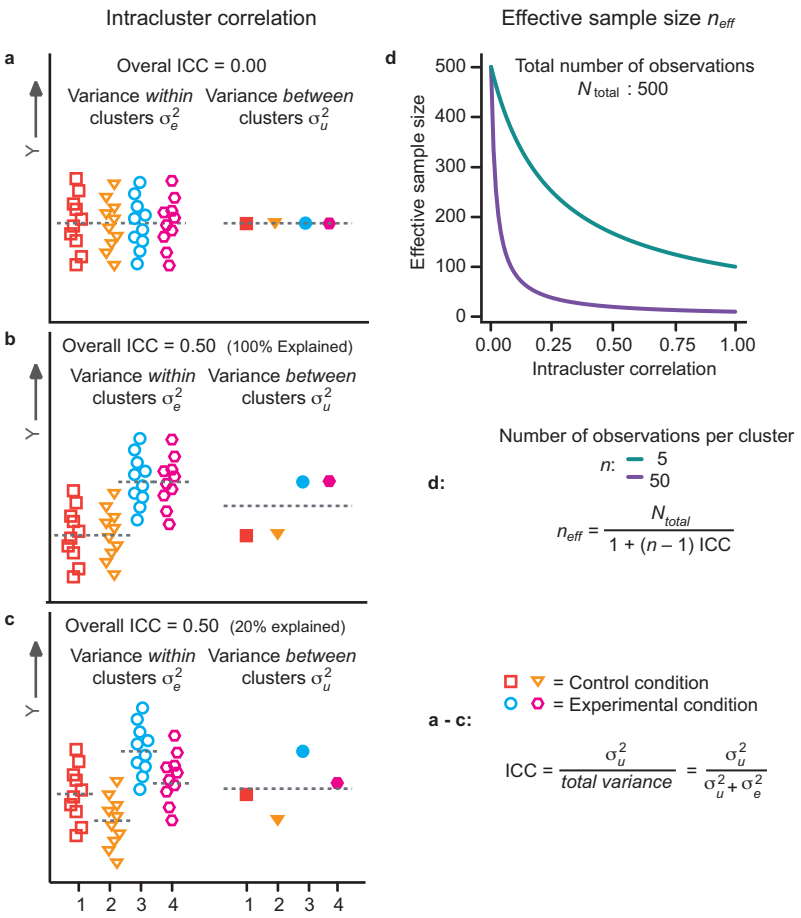

Fig. 2 Graphs illustrating why clustering affects the Type I error rate. (**a**–**c**). Graphical representation of three data sets with overall intracluster correlations ICC of (**a**) 0.00, (**b**) 0.50, fully explained by experimental condition and (**c**) 0.50, partly explained by experimental condition, respectively. The ICC is calculated from the variance between clusters (inferred from the deviations of the cluster means from the grand mean, represented by the grey horizontal line) and the total variance (i.e., the sum of the variance between clusters and the variance within clusters, calculated from the deviations of individual observations from their cluster mean). (**d**) Effective sample size as function of the ICC under two conditions (number of observations per clusters is 5 or 50, total number of observations is always 500). The higher the unexplained ICC, the larger the difference between the observed sample size ($N = 500$) and the effective sample size. The difference between the observed and effective sample size increases faster when the number of observations per cluster is higher.

Box 2: Inflation of the Type I error in nested data

By considering the error variance (i.e., the squared standard error, SE^2) of the experimental effect β_1, we show why conventional regression on nested data leads to inflated Type I error rate (i.e., probability of incorrectly rejecting the null-hypothesis). In multilevel analysis, the SE^2 of the experimental effect β_1 is

$$SE^2_{\beta_1} = \frac{n \times ICC + \sigma^2_e}{n \times N}, \quad (1)$$

where n represents the number of observations per clusters, ICC the unexplained intracluster correlation, N the number of clusters and σ^2_e the residual error variance (see Fig. 2 for an graphical representation of the individual statistical terms). In conventional regression (i.e., the t-test in regression terms), SE^2 is

$$SE^2_{\beta_1} = \frac{\sigma^2_e}{n \times N}. \quad (2)$$

Consequently, if clustering is not accommodated, the SE^2 is underestimated. The degree of underestimation depends on the number of observations per cluster n and the magnitude of the unexplained ICC (i.e., in equation 2, $n\times$ ICC is missing in the numerator). Note that in using conventional regression on clustered data, the residual error variance σ^2_e is actually a composite of the residual error variance and variance due to clustering. Also, note that equations 1 and 2 assume a standardized model (i.e., all variables have a mean of 0 and standard deviation of 1), a balanced design (i.e., the number of observations per cluster are equal and the number of clusters are evenly divided over the experimental groups), and absence of covariates.

2 The prevalence of nesting in Neuroscience studies

To assess the prevalence of nested data and the ensuing problem of inflated Type I error rate in neuroscience, we scrutinized all molecular, cellular and developmental neuroscience research articles published in five renowned journals (*Science*, *Nature*, *Cell*, *Nature Neuroscience*, and every month's first issue of *Neuron*) in 2012 and the first six months of 2013. Unfortunately, precise evaluation of the prevalence of nesting in the literature is hampered by incomplete reporting: not all studies report whether multiple measurements were taken from each research object, and if so, how many. Still, at least 53% of the 314 examined articles clearly concerned nested data, of which again 44% specifically reported the number of observations per cluster with a minimum of 5 observations per cluster (i.e., for robust multilevel analysis a minimum of 5 observations per cluster is required[11, 12], see Discussion). The median number of observations per cluster, as reported in literature, was 13 (Fig. 1a), yet conventional analysis methods were used in all these reports.

The studies reporting nested data typically do not provide information on the ICC, which is required to evaluate the extent to which clustering affected the Type I error rates of these studies. To assess the range of ICC's that can be expected, we analyzed 36 research questions in 18 neuroscience data sets from

varying disciplines. In these data, unexplained ICC's ranged between 0.00 and 0.74, with a mean of 0.19 (see Supplement 1.2 "Analysis results of 18 neuroscience datasets containing nesting" for full results). Since even a low degree of dependency (e.g., ICC = 0.10) increases the Type I error rate from 5% to nearly 20% when the number of observations per cluster is 13 (the median number of observations per cluster observed in our literature study; Fig. 1a), an excess number of false positive results is to be expected. Important to note is that differences in the statistical significance of the results between multilevel analysis and conventional testing are due to the unexplained ICC only, not the total ICC (see e.g., results of analysis 7, where all of the ICC is explained). Further inspection of the research articles that reported nested data with a minimum of 5 observations per cluster showed that 25% of the *p*-values were between .01 and .001, and 31% between .05 and .01. False positive effects are to be expected in at least some of these articles.

Moreover, another 7% of the examined papers used cluster-based summary statistics, where multilevel analysis could have been applied, resulting in a loss of power to detect experimental effects (see Fig. 1b and Supplement 1.2 "Analysis results of 18 neuroscience datasets containing nesting" for examples of non-significant results obtained with pooled t-tests, which actually prove significant when multilevel analysis is used).

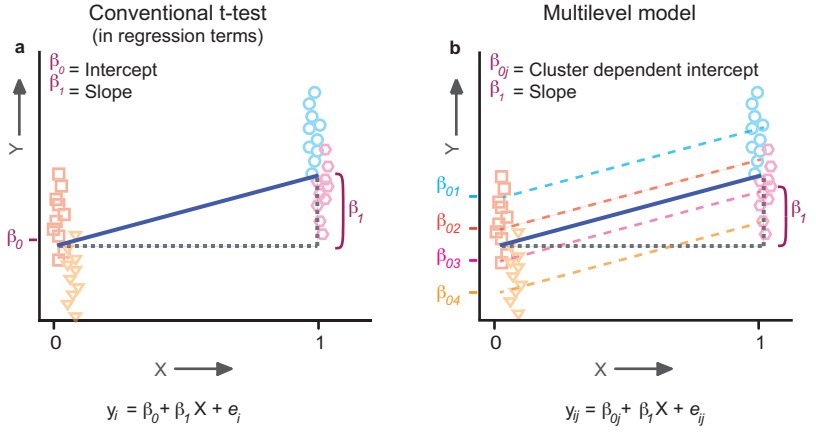

Fig. 3 Graphical representations of conventional *t*-test and multilevel analysis. (**a**) Graphical representation of the conventional *t*-test in regression terms: the individual observations y_i are a function of the mean of the control group (i.e., the intercept β_0), and when applicable, the estimated deviation from this mean for observations from the experimental group (i.e., the slope β_1), and an individual error term e_i. X is essentially a weight variable that takes on values 0 and 1 for observations from the control and experimental group, respectively. (**b**) Graphical representation of multilevel analysis. The individual outcomes of observation i from cluster j, y_{ij}, are a function of the cluster specific intercept β_{0j} plus, when applicable, the estimated deviation from this intercept for clusters belonging to the experimental group, β_1, and an individual-specific error term e_{ij}. The higher the unexplained ICC, the more variation there is in the cluster specific intercepts β_{0j}.

3 Multilevel analysis

Multilevel models can be used to statistically accommodate dependence between observations in nested designs. The basics of multilevel analysis are readily explained with reference to the conventional two-group t-test. Suppose we studied whether characteristic X of the cell is affected by a specific gene-mutation. In 15 mice carrying the mutation we collect 10 cells (i.e., 15 $\times$ 10 observations), and we do the same in 15 mice that do not carry the mutation, resulting in 300 observations in total. A standard t-test on these data can be carried out by regressing X on the dummy coded (0/1) experimental variable. Significance of the slope parameter, representing the differences in means, can be tested using a t-test (Fig. 3a). In this conventional analysis, cluster information is discarded: all 15 $\times$ 10 observations within each group are simply pooled. In contrast, in multilevel analysis the individual observations (here cells) are regarded as level 1 units, which are nested in the level 2 units: the clusters (here mice). Multilevel analysis retains cluster-membership information by conducting the t-test on the cluster-level (i.e., mouse) means, while retaining the distinction between the variance within clusters (i.e., differences between cells within a mouse) and variance between clusters (i.e., differences between the mice in cluster-level means; Fig. 3b). Multilevel analysis thus effectively accommodates the possibly increased similarity of observations taken from the same research object by retaining cluster-membership information of each individual observation, when evaluating parameters such as group differences.

Various standardized effect size measures have been suggested in the context of multilevel analysis [13, 14]. When comparing only two experimental conditions, Cohen's d is a generally accepted index. When comparing more than two experimental conditions, the overall effect size can be represented by the explained variance R^2, which equals the ICC when the experimental condition only varies over clusters and not within clusters. Cohen [15] defined a Cohen's d of 0.20, 0.50, and 0.80, and an explained variance R^2 of 0.01, 0.09, and 0.25 as small, medium and large effects, respectively. Note that these two effect sizes are not on the same scale and can therefore not be compared directly. However, d and R^2 can be converted into each other using these formulas [16]:

$$R^2 = \left(\frac{d}{\sqrt{d^2+4}}\right)^2 \text{ and} \tag{3}$$

$$d = \sqrt{\frac{4R^2}{1-R^2}}. \tag{4}$$

Note that in the first formula, the sum of the Cohen's ds obtained in pairwise comparisons is used when multiple pair-wise comparisons are combined into one omnibus test. Equations 3 and 4 assume experimental groups with equal sample sizes (see Lipsey and Wilson [16] for formulas for unbalanced designs).

A worked example of the analysis of nested data, including effect size calculation, is provided in Supplement 1.3 "A worked example of the analysis of nested data".

4 Power up: determining the optimal study design

Generally, power is increased by increasing the number of observations in a study. In conventional analysis, this is straightforward, but in multilevel analysis the relation between sample size and power is more complicated as the total number of observations is distributed over the research objects (i.e., clusters). In the allocation of research resources (e.g., money, time), a trade-off must be considered between the number of clusters and the number of observations per cluster. In practice, collecting many measures in a few clusters may be easier, faster and cheaper than collecting a few measures in many clusters. But which option confers the greatest power?

In multilevel analysis, power depends essentially on the number of clusters: power steadily increases to 100% as the number of clusters increases (Fig. 4a). In contrast, when increasing the number of observations per cluster, the power curve often approaches an asymptote below 100%, with the maximum level depending on the ICC (Fig. 4b). In general, high ICC's result in lower power, and unless the ICC is low, adding extra observations per cluster does little to increase power (Fig. 4a–b).

Given available resources (e.g., money, time), the optimal balance between the number of clusters (N) and number of observations per cluster (n) can be determined, given a specified level of dependency (ICC). In theory, optimal N and n are dictated by the desired level of statistical power. In practice,

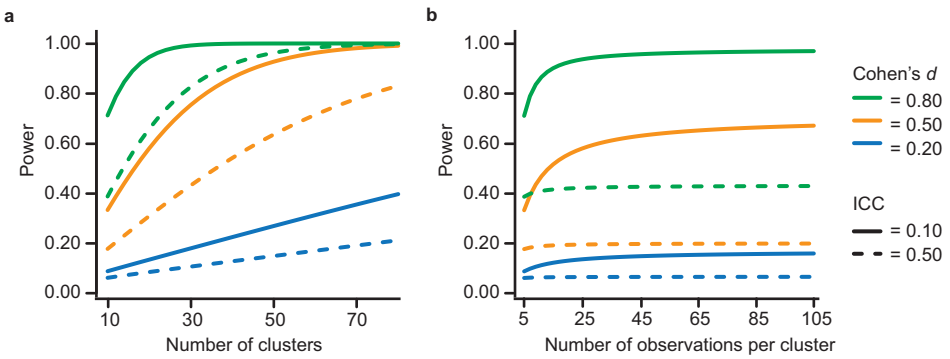

Fig. 4 Power of multilevel analysis to detect the experimental effect. Power is depicted under six conditions (Cohen's d of 0.20, 0.50 or 0.80, and unexplained intracluster correlation ICC of 0.10 or 0.50, respectively) and as function of (**a**) the number of clusters or (**b**) the number of observations per cluster. In (**a**), the number of observations per cluster is held constant at 5, in (**b**), the number of clusters is held constant at 10. Evidently, the number of clusters, and not the number of observations per cluster, is essential to increase the statistical power to detect the experimental effect.

however, available resources have a bearing on the attainable values of N and n. As including additional observations within a cluster (C_1) is usually less costly than including an additional cluster (C_2), these two costs are defined distinctly. The total costs of a study are calculated as

$$T = N \times (C_1 \times n + C_2), \tag{5}$$

while the optimal number of observations per cluster can be obtained by solving [9, 13]

$$n_{optimal} = \sqrt{\frac{C_2}{C_1} \times \frac{\sigma_e^2}{ICC}}, \tag{6}$$

where σ_e^2 is the residual variance, which equals 1− overall ICC (note that we make use of the standardized model, i.e., the observations are standardized such that they have a mean of 0 and a standard deviation of 1). Given the total available resources T, and the optimal number of observations per cluster $n_{optimal}$, the optimal number of clusters N can be obtained by

$$N = \frac{T}{n_{optimal} \times C_1 + C_2}. \tag{7}$$

The optimal balance between the number of clusters and number of observations per cluster does not guarantee that the subsequent study will have sufficient power to detect the experimental effect of interest. The actual power of the experiment additionally depends on the chosen alpha level and on the expected effect size (e.g., the magnitude of the difference between the control and experimental group). However, the calculated optimal N and n can be used to estimate the expected power given specific values of the effect size and the ICC: formulas and a worked example are provided in Box 3.

5 Discussion

Multilevel modeling is relevant to neuroscientific data collected using traditional techniques, such as the analysis of immunofluorescent signal intensity in slices (where the use of cluster-based summary statistics causes a loss of power), and the analysis of electrophysiological parameters, such as EPSPs (where the use of conventional statistical models inflates Type I error rates). Recent advances in the field of neuroscience, like optogenetics, super-resolution microscropy, immunogold cytochemistry, and optopharmacology, will, if anything, increase the relevance of multilevel modeling [17]. A common feature of all these novel techniques is that they shift the n from the animal or tissue level to the cellular or even subcellular level, and invariably yield data with a nested structure. For instance, super-resolution light microscopy allows imaging and advanced understanding of neuronal compartments[18], immunogold cytochemistry allows determination of subcellular localization of proteins[19],

Box 3: Estimating the power to detect an experimental effect

In this section, we discuss statistical power $(1-\beta)$ in the context of multilevel data, i.e., the probability of detecting an experimental effect that is actually present. Here the Type II error rate (β) is the probability of not rejecting the false null-hypothesis (i.e., in truth $\beta_1 \neq 0$). In multilevel analysis, the statistical significance of the experimental effect β_1 is tested by referring the Z statistic β_1/SE_{β_1} to the standard normal distribution. In this Z-test, the Z-statistic reflects the number of standard deviations that β_1 deviates from the expected value under the null-hypothesis (i.e., 0), from which a p-value for β_1 can be calculated. Power can be calculated by obtaining the estimated error variance (i.e., SE^2) of β_1, using the estimated error variance to convert β_1 to a Z-statistic, and subsequently obtaining the probability that the Z-score for β_1 exceeds the critical value for the noncentral Z-distribution given α. Below, we discuss the power calculation stepwise. When calculating power, it is easiest to work from the standardized model (i.e., both dependent and independent variable(s) have a mean of 0 and standard deviation of 1), because in that case, the difference in means between the experimental and control group equals the effect size Cohen's d, and the residual error σ_e^2 equals 1 - overall ICC. The equation to obtain the estimated error variance SE^2 of the experimental effect β_1 is given in equation 1. Next, as[12]

$$\frac{\beta_1}{\sqrt{SE^2_{\beta_1}}} = Z_{1-\alpha} + Z_{1-\beta}, \tag{8}$$

power can be estimated as

$$Z_{1-\beta} = \frac{\beta_1}{\sqrt{SE^2_{\beta_1}}} - Z_{1-\alpha} + , \tag{9}$$

The critical value for $Z_{1-\alpha}$ (i.e., the boundary value for which the null-hypothesis will be rejected) can be obtained from a Z-distribution table by locating the Z-statistic that corresponds to the value of $1-\alpha$. Note that for a two-sided test, $Z_{1-\alpha}$ needs to be substituted by $Z_{1-\alpha/2}$ in equations 8 and 9. For instance, for a two-sided test with $\alpha = 0.05$, $Z_{1-\alpha/2}$ equals 1.96. The probability of the outcome value $Z_{1-\beta}$ can be obtained from a Z-distribution table by locating the probability that corresponds to the Z-statistic . Note that when using a standardized model, the experimental effect β_1 is half of the difference between the control and experimental group (i.e., in the standardized model assuming equal group sizes, the experimental variable X is coded as -1 and 1 instead of 0 and 1).

To illustrate, suppose we are planning a study on the differences between wild type and knockout mice with respect to a cell characteristic in primary cultures. We are planning to use 64 clusters (e.g., primary cultures) with 12 observations per cluster in total (e.g., the optimal number of clusters and observations per cluster when we would have 4,000 monetary units to spend, the costs of plating a primary culture are 50 monetary units and the costs to obtain an observation from one cell of this primary culture equals 1 monetary unit, see equations 5–7). Based on previous data, we assume that the unexplained ICC is approximately 0.25. As the effect size is unknown, we obtain an estimate of the power to detect a small ($d = .2$) and a medium ($d = .5$) difference between genotypes. Using equation 3, the difference between genotypes relates to an explained ICC of .01 and .06, respectively. Accordingly, σ_e^2 is set to $1 - 0.25 - 0.01 = 0.74$ and $1 - 0.25 - 0.06 = 0.69$, respectively.

Box 3: Estimating the power to detect an experimental effect - continued

Since β_1 is calculated as $d \times .5$, the β_1 for the small and medium effects correspond to $\beta_1 = .2 \times .5 = .1$ and $\beta_1 = .5 \times .5 = .25$, respectively.
The power calculations assuming a two-sided test with $\alpha = 0.05$ are as follows. The estimated error variance SE^2 for the experimental effect equals $(12 \times 0.25 + 0.74)/(12 \times 64) = 0.005$ and $(12 \times 0.25 + 0.69)/(12 \times 64) = 0.005$ for a small and medium difference between genotypes. For a small difference between genotypes, $Z_{1-\beta}$ equals $(0.1/\sqrt{0.005}) - 1.96 = -0.546$. The probability of $Z_{1-\beta}$, obtained using the Z-distribution table, is 0.29, so we have an estimated power of 29%.
Along the same lines, the power for detecting a medium difference between genotypes equals 94%. We conclude that the estimated power to detect a small experimental effect is too low. If we want a larger probability to detect a small experimental effect, more resources are needed in order to increase our sample size. Given that the cost ratio and the ICC stay equal, the optimal number of observations per cell remains 12 (see equation 6). Therefore, we only need to calculate how many extra cells we can afford given increased resources. If we tripled our resources to 12,000 monetary units, we could triple the number of primary cultures, which comes to 195 platings (see equation 7). The power to detect a small experimental effect now increases to 69%. If we are only interested in detecting a medium effect size, our initial resources certainly suffice (given the calculated power of 94%

and recent advances in optogenetics and optopharmacology facilitate selective control of respectively electrical and protein activity in circuits, individual cells, or subcellular compartments[20, 21]. All these techniques concern the collection of multiple observations from one cell, and thus yield nested data.
To fully exploit the advantages that these novel techniques offer, neuroscientist should adopt multilevel modeling to avoid the limitations of conventional analyses in this context. In addition, nested data come with specific design issues that are relevant to the statistical power to resolve the effects of interest. Optimalization of design in terms of allocation of resources does not guarantee sufficiently powered studies. In terms of power, the ratio of number of research objects (e.g., mice) to the number of measurements per object (e.g., cells per mouse) is important. We showed that the power increase achievable by increasing the latter is limited (Fig. 4). In addition, to obtain robust and unbiased estimates of variance components in multilevel analysis, sufficient observations on both levels are required. As a rule of thumb, afforded by simulation studies[11, 12], minimally 5 observations per cluster, and 10 clusters per experimental group are recommended to obtain a robust and unbiased estimate of the standard error for the experimental effect. To also obtain a robust and unbiased estimate of the intracluster correlation, the number of clusters needs to be increased to 30.
Here we focused on the most common design, i.e., data that span two levels (e.g., cells in mice) and an experimental variable that does not vary within clusters (e.g., in comparing cell characteristic X between mutants and WTs, all cells from one mouse have the same genotype). Other nested designs –

featuring three or more levels of nesting, experimental variables that do vary within levels (e.g., when investigating whether the number of docked vesicles differs between observations from a dendrite or an axon), nested longitudinal data (i.e., data collected on multiple time points describing dynamical processes[22, 23]), or nested non-normally distributed data (e.g., binary or Poisson distributed data) – are, however, possible and can be analyzed using multilevel analysis. We refer to [12, 14, 24] for comprehensive introductions to multilevel modeling, and to the Centre for Multilevel Modeling website (http://www.bristol.ac.uk/cmm/learning/mmsoftware/) for a recent overview of existing multilevel software.

Various recent publications force neuroscientists to acknowledge the possibility that the harvest of their hard labor is contaminated by an abundance of false positive effects[1–4]. Nested designs are ubiquitous in neuroscience, and increased awareness of the problem of nesting in both researchers and reviewers will prevent costly and time-consuming quixotic pursuits of spurious effects, and thus assist progress in the understanding of the nervous system.

Acknowledgments

We are very grateful to our colleagues from the VU University/VU Medical Center Functional Genomics department for sharing their data. M.V. is supportedby the European Union (ERC Advanced grant 322966; HEALTH–F2–2009–241498 EUROSPIN, and HEALTH–F2– 2009–242167 SynSys) and the Netherlands Organization for Scientific Research (TOP 903–42–095). C.V.D. is supported by the European Research Council (Genetics of Mental Illness, grant number: ERC–230374). S.v.d.S. is supported by the Netherlands Scientific Organization (Nederlandse Organisatie voor Wetenschappelijk Onderzoek, gebied Maatschappij- en Gedragswetenschappen: NWO/MaGW: VIDI–452–12–014).

COMPETING FINANCIAL INTERESTS

The authors declare no competing financial interests.

References

[1] Button, K.S., Ioannidis, J.P., Mokrysz, C., Nosek, B.A., Flint, J., Robinson, E.S., Munafò, M.R.: Power failure: why small sample size undermines the reliability of neuroscience. Nature Reviews Neuroscience **14**(5), 365–376 (2013)

[2] Ioannidis, J.P.: Why most published research findings are false. Chance **18**(4), 40–47 (2005)

[3] Nieuwenhuis, S., Forstmann, B.U., Wagenmakers, E.-J.: Erroneous analyses of interactions in neuroscience: a problem of significance. Nature neuroscience **14**(9), 1105–1107 (2011)

[4] Tsilidis, K.K., Panagiotou, O.A., Sena, E.S., Aretouli, E., Evangelou, E., Howells, D.W., Salman, R., Macleod, M.R., Ioannidis, J.P.: Evaluation of

excess significance bias in animal studies of neurological diseases. PLoS Biol **11**(7), 1001609 (2013)

[5] Raising standards. Nature Neuroscience **16**(5), 517–517 (2013)

[6] Making methods clearer. Nature Neuroscience **16**, 1 (2013)

[7] Galbraith, S., Daniel, J.A., Vissel, B.: A study of clustered data and approaches to its analysis. J Neurosci **30**(32), 10601–8 (2010)

[8] Walsh, J.E.: Concerning the effect of intraclass correlation on certain significance tests. The Annals of Mathematical Statistics, 88–96 (1947)

[9] Raudenbush, S.W.: Statistical analysis and optimal design for cluster randomized trials. Psychological Methods **2**(2), 173 (1997)

[10] Moerbeek, M., van Breukelen, G.J., Berger, M.P.: A comparison between traditional methods and multilevel regression for the analysis of multicenter intervention studies. Journal of clinical epidemiology **56**(4), 341–350 (2003)

[11] Maas, C.J., Hox, J.J.: Robustness issues in multilevel regression analysis. Statistica Neerlandica **58**(2), 127–137 (2004)

[12] Snijders, T., Bosker, R.: Multilevel Analysis: An Introduction to Basic and Applied Multilevel Analysis. Sage, London (2011)

[13] Snijders, T.A., Bosker, R.J.: Standard errors and sample sizes for two-level research. Journal of Educational and Behavioral Statistics **18**(3), 237–259 (1993)

[14] Hox, J.J., Moerbeek, M., van de Schoot, R.: Multilevel Analysis: Techniques and Applications. 2nd Edn. Routledge, New York, NY (2010)

[15] Cohen, J.: Statistical Power Analysis for the Behavioral Sciences. 2nd Edn. Erlbaum, Hillsdale, New Jersey (1988)

[16] Lipsey, M.W., Wilson, D.B.: Practical Meta-analysis. Sage Publications, Inc, London (2001)

[17] Focus on neurotechniques. Nature Neuroscience **16**(7), 771 (2013)

[18] Maglione, M., Sigrist, S.J.: Seeing the forest tree by tree: super-resolution light microscopy meets the neurosciences. Nature neuroscience **16**(7), 790–797 (2013)

[19] Amiry-Moghaddam, M., Ottersen, O.P.: Immunogold cytochemistry in neuroscience. Nature neuroscience **16**(7), 798–804 (2013)

[20] Packer, A.M., Roska, B., Häusser, M.: Targeting neurons and photons for optogenetics. Nature neuroscience **16**(7), 805–815 (2013)

[21] Kramer, R.H., Mourot, A., Adesnik, H.: Optogenetic pharmacology for control of native neuronal signaling proteins. Nature neuroscience **16**(7), 816–823 (2013)

[22] Smith, A.C., Stefani, M.R., Moghaddam, B., Brown, E.N.: Analysis and design of behavioral experiments to characterize population learning. Journal of Neurophysiology **93**(3), 1776–1792 (2005)

[23] Czanner, G., Eden, U.T., Wirth, S., Yanike, M., Suzuki, W.A., Brown, E.N.: Analysis of between-trial and within-trial neural spiking dynamics. Journal of neurophysiology **99**(5), 2672–2693 (2008)

[24] Goldstein, H.: Multilevel Statistical Models, 4th edn. John Wiley & Sons, West Sussex (2011)