Toolbox

Editor's Note: Toolboxes are intended to describe and evaluate methods that are becoming widely relevant to the neuroscience community or to provide a critical analysis of established techniques. For more information, see http://www.jneurosci.org/misc/ifa_minireviews.dtl.

A Study of Clustered Data and Approaches to Its Analysis

Sally Galbraith,[1]* James A. Daniel,[2,3]* and Bryce Vissel[2,4]

[1]School of Mathematics and Statistics, University of New South Wales, Sydney 2052, Australia, [2]Neuroscience Program, Garvan Institute of Medical Research, Sydney 2010, Australia, [3]Cell Signalling Unit, Children's Medical Research Institute, Sydney 2145, Australia, and [4]St Vincent's Medical School, University of New South Wales, Sydney 2052, Australia

Statistical analysis is critical in the interpretation of experimental data across the life sciences, including neuroscience. The nature of the data collected has a critical role in determining the best statistical approach to take. One particularly prevalent type of data is referred to as "clustered data." Clustered data are characterized as data that can be classified into a number of distinct groups or "clusters" within a particular study. Clustered data arise most commonly in neuroscience when data are compiled across multiple experiments, for example in electrophysiological or optical recordings taken from synaptic terminals, with each experiment providing a distinct cluster of data. However, there are many other types of experimental design that can yield clustered data. Here, we provide a statistical model for intracluster correlation and systematically investigate a range of methods for analyzing clustered data. Our analysis reveals that it is critical to take data clustering into account and suggests appropriate statistical approaches that can be used to account for data clustering.

Introduction

Statistical analysis is a foundation of all biological research. Various methods exist to test whether there is a difference in the response, for example, between a "control" and a "treated" sample. Observations are generally made across a series of independent experiments to ensure that a result is reproducible. Results from a single experiment (such as reporting measurements from a single electron micrograph) lack reproducibility. Statistically significant differences between groups (typically, $p < 0.05$) indicate that the observed difference is unlikely to have occurred by chance, suggesting a "real" difference between groups. Research outcomes rely on the presentation of statistically valid conclusions, and thus the approach used for statistical analysis is critical.

Statistical analysis must consider features of the data, including the measurement scale (e.g., continuous, binary, or categorical), the experimental units, and the way in which the data were collected. One type of data that arises from certain data collection schemes or from the way the experimental units are structured is clustered data.

Clustered data are frequently obtained in the neurosciences, but rarely is its analysis discussed explicitly in neuroscience literature. There are many useful references throughout the statistical literature that discuss clustered data (for examples, see Brown and Prescott, 1999; Gonen et al., 2001; Zyzanski et al., 2004). The aim of our paper is to consolidate some of the issues surrounding clustered data, present them in a form accessible to neuroscientists, and provide ways of dealing with what is a widely encountered type of data in neuroscience. We undertake a study of clustered data and demonstrate that it is critical to take clustering into account in the analysis of data generally in neuroscience. We address this in two parts. In Part 1, we will describe the nature of clustered data and provide examples of approaches to analyze clustered data. In Part 2, we study the effects of using these different statistical approaches in the analysis of clustered data. We conclude with suggestions on the most effective methods for dealing with clustered data. As we will show, using an appropriate statistical approach that takes clustering into account may critically impact the results of an analysis and, hence, the conclusions of a study.

Part 1: What is clustered data?

The term "clustering," as used in this paper, is not related to the statistical technique "cluster analysis," which is an unsupervised learning technique used to uncover hidden structure in the data. Instead, clustering will be apparent from the way the data are collected, as discussed below.

Received Jan. 18, 2010; revised March 3, 2010; accepted May 14, 2010.

Our work was supported by National Health and Medical Research Council Australia Grant 188819 (B.V.), a New South Wales (NSW) State Government Spinal Cord Injury and Related Neurological Conditions Research Grant administered by the Office for Science and Medical Research (B.V.), the NSW State Government's BioFirst Award (B.V.), the Australian Postgraduate Award (J.A.D.), the Baxter Family Scholarship (J.A.D.), the Gowrie Trust Scholarship (J.A.D.), Amadeus Energy Ltd., Bill Gruy, Geoff and Dawn Dixon, Walter and Edith Sheldon, and the Henry H. Roth Charitable Foundation. This work would not have happened without Amadeus Energy Ltd., Bill and Laura Gruy, Geoffrey Towner, Walter and Edith Sheldon, and Tony and Vivian Howland Rose. We thank Nick Kell and Joanna Knott for their support.

*S.G. and J.A.D. contributed equally to this work.

Correspondence should be addressed to Dr. Bryce Vissel, Neuroscience Program, Level 7, Garvan Institute of Medical Research, 384 Victoria Street, Darlinghurst, North South Wales 2010, Australia. E-mail: b.vissel@garvan.org.au.

DOI:10.1523/JNEUROSCI.0362-10.2010

 0270-6474/10/3010601-08$15.00/0

Clustered data arise when the data from the whole study can be classified into a number of different groups, referred to as clusters. Each cluster contains multiple observations, giving the data a "nested" or "hierarchical" structure, with individual observations nested within the cluster. The key feature of clustered data is that observations within a cluster are "more alike" than observations from different clusters.

There are many examples where clustered data occurs in neuroscience, such as the following:

(1) Studies that obtain data from multiple experiments where several observations are collected from each experiment. Consider, for example, a study of synapse populations comprised of a series of independent experiments where each experiment yielded data from multiple synapses. In this example, all data obtained from the same experiment, i.e., synapse data obtained from the same section, synapse data obtained from the same slide, or synapse data obtained from the same animal, constitute a cluster.
(2) Longitudinal data, where multiple measurements are taken over time on each individual. For example, in animal models of neurological disease the measurements at different times on the same animal form a cluster.
(3) Multicenter clinical trials, where a cluster consists of measurements on patients from the same center.
(4) Cluster randomized trials where, for example, whole clinics are randomized to an intervention. Here, the clusters are formed of patients within clinic.
(5) Genetic epidemiology studies using family data. Here, data on members of the same family constitute a cluster.

The fact that observations within a cluster are more alike than observations from different clusters induces a correlation between observations within the same cluster. This is referred to as intracluster correlation. Thus, observations within a cluster are correlated, whereas observations from separate clusters are regarded as independent. Since observations within a cluster do not contribute completely independent information, the "effective" sample size is less than the total number of observations from all clusters.

Part 1.1: A specific example of clustered data

Our focus is on testing hypotheses by comparing observations from two groups, such as a treated group versus a control group. As a case in point, consider a hypothetical experiment directed to assess whether a drug leads to altered rates of exocytosis from synapses *in vitro*. To address this question on a given day, neurons on 10 coverslips are each to be treated with 10 μg/ml drug, while neurons on another 9 coverslips are to be treated with vehicle. The following day the rate of exocytosis at between 10 and 40 synapses per coverslip is to be determined. Let us say that such an experiment yielded exocytosis rate data on a total of 123 synapses from 10 drug-treated coverslips and data from 157 synapses from 9 untreated coverslips. The aim of the study is to test the null hypothesis that there is no effect of the drug on the exocytosis rate.

In fact, in this hypothetical experiment, the synapses are nested within coverslip and, hence, the data obtained from individual coverslips constitute clusters. For each synapse, we observe unique exocytosis kinetics, and the exocytosis rate can be estimated. The exocytosis rates for treated and untreated synapses could be compared using, for example, a two-sample *t* test (if the data are approximately normally distributed). However, it is possible that other conditions within a coverslip, apart from treatment, are more similar than conditions on different coverslips. Hence, the synapses from the same coverslip share similar conditions and may be more or less likely to release neurotransmitter through exocytosis than synapses from different coverslips, quite apart from the treatment effect.

In the above example, the individual observations belong to a number of different clusters. While we are not usually primarily interested in the cluster effect, it must be taken into account to ensure validity of the treatment comparison. If the clustering is not taken into account for this type of data, then the variability is likely to be underestimated and the results of the analysis invalidated.

Part 1.2: Two classes of clustered data

Throughout this document, we will refer to two classes of clustered data, which for convenience we will call case 1 data and case 2 data.

In the hypothesis testing context, it is useful to distinguish between two different kinds of clustered datasets: case 1, where only one of the groups that is being compared is represented in each cluster; and case 2, where (at least some of) the clusters contain observations from both groups. The hypothetical experiment in Part 1.1 above is an example of case 1. Another example of case 1 is a cluster randomized trial where, for practical reasons, it is necessary to randomize all of the patients from a particular clinic to the same intervention. For example, all patients from one clinic might receive the usual standard of care, while all patients from another clinic participate in an additional health care education program. Here, the unit of randomization is the clinic. The data from each clinic is clustered and each cluster will contain either patients receiving standard care or patients receiving the education program in addition to standard care. An example of case 2 is a multicenter clinical trial where, to widen the patient base, patients are recruited from a number of different centers, but within each center some patients will be randomized to one treatment and some to the other treatment. Here, the unit of randomization is the patient. Data from each clinic are clustered, but this time the clusters will contain both control and treated patients. The best methods for analyzing case 1 and case 2 data may differ, and this will be discussed in later sections.

It is a fact that clustered data are very common in neuroscience. Experiments yield multiple observations, animals yield observations on multiple litter members, or, at a series of times, research centers contribute data on multiple patients. In each case we are faced with clustered data that raise the question: is conventional statistical analysis sufficient, or should a method be used that specifically accounts for clustering? We shall address this question in detail.

Part 1.3: A statistical model for intracluster correlation

Suppose we collect data from several independent experiments that represent the clusters. The similarity between clusters might be accounted for by assuming that there is a common mean response within a cluster that varies randomly between clusters. The response within clusters also varies due to individual-level heterogeneity. This leads to the model for the data, $y_{ik} = \mu + b_k + \varepsilon_{ik}$, where y_{ik} is the value of the response variable for unit i in cluster k, and μ is the overall mean. The remaining two terms represent the two levels of variation in the data, with ε_{ik} representing the "within-cluster" variation between observations from the same cluster, and b_k the

"between cluster" variation. In statistical terms, b_k is a random cluster-specific effect with mean zero and variance σ_b^2, and ε_{ik} is a residual error term with mean zero and variance σ_w^2. The residual errors from different individuals (ε_{ik}) and the random effects from different clusters (b_k) are assumed to be independent. In this model, the correlation between two different observations from the same cluster arises because of the random effect b_k, and this correlation is equal to $\sigma_b^2/(\sigma_b^2 + \sigma_w^2)$. While observations within each cluster are correlated, there is no correlation between observations from different clusters because the random effects from different clusters and the residual errors from different individuals are independent.

Part 1.4: Analysis of clustered data

Having defined clustered data, we will now address the various ways in which clustering can be treated. In reviewing the literature, it would appear that four approaches have generally been used in the analysis of clustered data: (A) ignoring clustering; (B) reducing clusters to independent observations; (C) fixed effects regression/ANOVA approaches; and (D) explicitly accounting for clustering.

Each of these approaches will be discussed in detail here. By way of example, we will make specific reference to optical analysis of presynaptic function. Optical methods of studying presynaptic function use fluorescent tracers, such as styryl dyes or synaptopHluorin, to mark synaptic vesicles that are undergoing release. Optical methods yield clustered data because each independent experiment provides data from multiple synapses that can be visualized individually under the microscope. Observations from a population of synapses within the same experiment are clustered and hence are more similar than observations of synapses acquired in separate experiments. Studies using these optical methods have adopted various approaches to analyze their data, and thus these studies provide an insight into ways in which clustered data may be approached.

Approach A: Ignoring clustering. Two questions that researchers may ask are: (1) do I have clustered data; and (2) should I use a method that accounts for clustering? Question 1 can be answered without any recourse to statistical tests; it is simply a matter of study design. If a study collects multiple observations in a number of different groups, then the data are clustered. Question 2 is really asking what the consequences of ignoring the clustering would be.

In the past, methods for handling clustered data have not been as well developed or widely understood as methods for independent data. Therefore, in many studies that generate clustered data the simplest approach has been adopted, namely to ignore the clustering and treat the data as if all observations were independent. The consequences of adopting this approach will depend on the nature of the correlation that actually exists in the data, which will be discussed in Part 2.

Many studies that involve optical recording from synapses have ignored clustering. In these studies, the data obtained from individual synapses are pooled to create a single large dataset. The studies then use the number of individual boutons analyzed as the statistical *n* value, that is, they assume the data gathered at each synapse to be an independent measurement. This assumption is not valid, as the studies are conducted across multiple experiments and therefore yield data that are clustered (i.e., the individual observations are not independent). It is possible that had a statistical method that accounted for clustering been used, the conclusions of some studies may potentially have been different.

Approach B: Reducing data to independent observations. Another widely used approach consists of two stages. The first stage is to reduce the multiple observations in a cluster to a single observation by taking a suitable summary, which is commonly the mean of all the observations in each cluster. The resulting data points are all from different clusters and are thus regarded as independent. The reduced data can then be analyzed using standard methods for independent observations (such as a *t* test). This approach is common in many types of study, for example in electrophysiology studies in which multiple recordings from a single cell are all averaged to give a single measurement.

This approach has been used by several studies that performed optical recording for some (Pyle et al., 2000; Mozhayeva et al., 2002; Krueger et al., 2003; Fernandez-Alfonso and Ryan, 2004) or all (Virmani et al., 2006; Willeumier et al., 2006) aspects of their analyses. The studies examined individual synapses but the data from each experiment were averaged, thereby reducing multiple observations to a single observation for each experiment.

Reducing the data to independent observations is a valid approach to studying clustered data. The correlation associated with data clustering has effectively been removed and no longer needs to be accounted for. Studies that reduce data to the cluster means need only assume that separate experiments are independent.

However, there are limitations to this approach. Firstly, if there are unequal numbers of observations per cluster, then an unweighted method of second stage analysis may not be the most appropriate. Simply taking the mean of each cluster and then comparing these values by a *t* test, for example, does not take the unequal number of observations per cluster into account. Clusters with more observations could be expected to contribute more information and, thus, should be given more weight in the analysis.

Another disadvantage is that by taking the average of the observations in each cluster, information regarding the individual observations is lost in the process. Reducing a cluster to its mean value is a trade off in which clustering is eliminated but information about the individual observations is lost. The loss of information may have the result that an analysis based on cluster means is less powerful than an approach that incorporates information on the individual observations. A number of studies have examined functional heterogeneity across populations of individual synapses (Murthy et al., 1997; Ryan et al., 1997; Murthy and Stevens, 1998; Waters and Smith, 2002; Moulder et al., 2007; Daniel et al., 2009). In these studies the examination of individual synapses was key and, hence, these studies could not have been performed if data from each independent experiment (i.e., cluster) were averaged.

A complication arises in the analysis of case 2 data (described in Part 1.2) in which a single cluster can contain observations from both of the groups that the investigators wish to compare. In this case, the data cannot be reduced to just a single measurement of the response for each cluster, since separate results need to be kept for the two groups being compared. The data could be reduced to a pair of observations per cluster, namely the mean response for each group. A paired comparison could then be performed, such as a paired *t* test for normally distributed data or a Wilcoxon signed rank test for non-normal data.

In sum, although reducing the observations in each cluster is strictly a valid approach to analyzing clustered data, there are cases in which simply taking the mean from each cluster will not be the best approach. It is with this in mind that we move on to the third and fourth approaches.

Approach C: Fixed effects regression/ANOVA approaches. The idea here is that the cluster effect could be taken into account by including it as a factor in a standard regression model. Hence, if the aim is to compare two groups, then we would set up a regression model with explanatory variables consisting of an indicator for "group" and, in the case of *K* clusters, (*K-1*) indicator variables for the clusters. For a normally distributed outcome, this is equivalent to an ANOVA approach.

To consider this approach in more detail, it is instructive to consider our earlier example regarding the comparison of exocytosis rates between treated and untreated synapses. Each coverslip receives only one treatment, either toxin or vehicle, and so each coverslip contains observations from only one treatment group. Thus, only one of the groups being compared is represented in each cluster, which we refer to in Part 1.2 as case 1 data. For case 1 data there is no within-cluster comparison of the groups available and, hence, insufficient information to estimate both the group effect and a fixed effect for each cluster. Thus, a fixed effects regression could not be used in this case. This is true for all case 1 data, regardless of whether data are normally distributed, skewed, binomial, etc. For our specific example, if the data follow a normal distribution, a valid method of analysis would be to use a linear mixed model, as discussed in approach D.

In studies that yield case 2 data (Part 1.2) a fixed effects regression approach may be suitable. Each cluster contains observations from both groups being compared, which allows a within-cluster comparison of the two groups. A model including both the group effect and a fixed effect for each cluster can be fitted. This analysis essentially "controls for" the cluster effect, which can be regarded as a nuisance, and estimates the group effect at a fixed level of cluster (that is, within each cluster).

However, there are some remaining issues with this approach. One issue is whether a fixed or random effect for cluster is more appropriate. If a fixed effect is used, then the results of the analysis are strictly only applicable to the particular set of clusters in the study. With a random cluster effect (as would be the case with a mixed model, discussed in approach D), the clusters are regarded as a random sample from a wider population of clusters, so the results can be generalized to the wider population. Similarly if a cluster by group interaction is to be fit, it may be conceptually more appealing to regard the group effect as varying randomly over the clusters. This approach also makes it possible to include an interaction effect even if some clusters only contain observations from one group; this would not be possible with a fixed effects approach. Additionally, if we are interested in the group comparison within a given cluster, then the random effects approach allows us to incorporate information from all of the clusters, whereas the fixed effects approach uses information from that cluster alone.

For certain types of "clustered" data a random effects approach is inherently much more appealing. For example, with longitudinal data where there are multiple measurements over time for each individual and we want to model a linear trend over time, incorporating a fixed effect for each individual would be very cumbersome and the random effects approach is almost universally used.

Finally, there may be situations where the correlation structure is itself of interest (for example, genetic studies). Use of a fixed effects model would not allow this correlation structure to be studied.

Nevertheless, use of a fixed effect for cluster is often possible for case 2 data and can offer benefits in terms of lower SEs, particularly when the number of clusters is small. The issues regarding the choice of fixed versus random effects are complex. For further discussion see, for example, the book by Brown and Prescott (1999) or the article by Senn (1998).

A final issue (that is not unique to the fixed effects regression approach) relates to the type of outcome variable. If the data are normal or can be transformed to normality, then a normal regression (ANOVA) approach with a fixed effect for cluster and an effect for group can be used. For nonnormal data, a generalized linear model could be used or, alternatively, a test such as the Wilcoxon rank sum test, suitably modified to account for clustering, could be used. A discussion of such tests appears in the next subsection.

Approach D: Methods that explicitly account for clustering. Several methods are available that explicitly account for the within-cluster correlation and, as such, are ideal for analyzing clustered data. The methods can be classified into two broad groups, which we detail below.

Group 1: Methods that adjust existing tests to account for clustering. These approaches are used to test the null hypothesis of no difference between two groups of observations.

The statistical test used depends very much on data distribution. A test for normally distributed, clustered data adjusts the standard two-sample *t* test by an additional factor designed to take account of the intracluster correlation. Thus, the *t* test can be modified to account for data clustering. A similar approach has been proposed for binary response data, in this case adjusting the usual χ^2 statistic (Donner and Banting, 1988). These methods have been reviewed in detail previously (Gonen et al., 2001).

When analyzing clustered data that are not normally distributed, rank-based tests have been developed. For case 1 clustered data, modifications of the Wilcoxon rank sum test (equivalently, the Mann–Whitney *U* test) have been proposed (Rosner and Grove, 1999; Rosner et al., 2003).

For case 2 data, in which a single cluster contains observations from both groups under comparison, rank-sum tests have also been developed (Datta and Satten, 2005; Rosner et al., 2006b; Larocque et al., 2010). To illustrate the application of such tests, we refer the reader to our recent study of dopaminergic synapses (Daniel et al., 2009). We identified two distinct types of bouton in each experiment: (1) synaptic boutons, which constituted conventional synapses consisting of both presynaptic and postsynaptic components; and (2) nonsynaptic boutons, which possessed no postsynaptic component. We used the Datta and Satten (2005) test to compare the probability of vesicle release at these two bouton populations. This approach allowed us to account for the non-normal distribution of the data, the cluster effect of having a dataset comprised of multiple independent experiments, and the fact that observations from each of the groups being compared were present in each experiment.

Signed-rank tests have also been developed for paired data, where the pairs are clustered (Rosner et al., 2006a,b; Datta and Satten, 2008).

Group 2: *Modeling approaches*. Modeling approaches are generally more complex than the methods discussed above but are particularly useful when there are other covariates that need to be included in the analysis, such as temperature or time of day. For normally distributed data, a linear mixed model (LMM) can be used (Laird and Ware, 1982). These models incorporate random cluster-specific effects, as in the statistical model for intracluster correlation described previously, which induce the within-cluster correlation. Generalized linear mixed models, an extension of LMMs, can be used for con-

Table 1. Parameters used in generating simulated data

Model number	Data distribution	Case	β	Number of clusters	Average cluster size	Range of cluster sizes	Number of observations
1	Normal		0	20	10	5–16	200
2	Normal	1	1	20	10	5–16	200
		2	0.2				
3	Skewed		0	20	10	5–16	200
4	Skewed	1	1	20	10	5–16	200
		2	0.3				
5	Normal		0	8	6	3–9	48
6	Normal	1	2	8	6	3–9	48
		2	0.5				
7	Skewed		0	8	6	3–9	48
8	Skewed	1	2.5	8	6	3–9	48
		2	1				

The simulations with $\beta = 0$ (no difference between group 1 and group 2 data) investigate how liberal/conservative the tests are (performance under the null), whereas the simulations with $\beta \neq 0$ (real difference between group 1 and group 2 data) investigate the power of the test (performance under the specified alternative). Case 1 refers to data where only a single group is represented in each cluster, and case 2 refers to data with approximately equal numbers from both groups in each cluster. For each of the 8 models, 10,000 datasets were generated.

tinuous non-normal, binary, and categorical responses. For survival data, random "frailties" can be used to induce within-cluster correlation in a manner akin to that used for mixed models.

Generalized estimating equations (GEEs) constitute another approach (Liang and Zeger, 1986), which can also be used for continuous non-normal, binary, and categorical responses. GEEs separately model the mean response and the within-cluster association, assuming primary interest is in the former and regarding the latter as a nuisance that must be taken into account for valid inference.

Thus, there are a number of analytical approaches that adequately account for intracluster correlation. These approaches are ideally suited to the analysis of clustered data. Moreover, the method used to analyze clustered data can have significant effects on results, as we will demonstrate in the next section.

Part 2: The effect of clustering on the outcomes of statistical analyses

The impact of clustering on the various methods at our disposal depends crucially on the strength of the intracluster correlation. Specifically, if the intracluster correlation is relatively strong, then the failure to take clustering into account is likely to have a more profound effect on the outcomes of statistical analyses than if it is weak, as we shall show. However, the impact of clustering on statistical analysis also depends on the distribution of the data (for example, symmetric or skewed), the number of clusters, and the number of observations per cluster.

In this section, we will use randomly generated datasets to illustrate how various methods of statistical analysis perform when analyzing clustered data. Randomly generated datasets, created from a known statistical model, enable the performance of the analytical methods to be compared with desired levels. This allows us to assess the suitability of each test with different types of data and different population sizes. We will manipulate several variables within the simulation so as to examine as many different scenarios as possible, thereby providing a comprehensive overview of how different kinds of statistical tests perform for different types of data. For interested readers, we also provide some mathematical intuition describing the effect of clustering on hypothesis testing (see supplemental Fig. S1, available at www.jneurosci.org as supplemental material).

First, we will examine how liberal/conservative a given statistical test is. To examine this question, we generate datasets according to a model that specifies no difference between the two groups being compared. We then test the null hypothesis of no difference between the two groups using a 5% significance level. Accordingly, we should see the null hypothesis rejected for about 5% of the datasets. A test that rejects the null hypothesis >5% of the time is too liberal, and one that rejects it <5% of the time is too conservative.

Second, we will investigate the power of a given statistical test, which indicates how effective a test is in detecting that a difference exists between two groups. To do this, we will generate data under an alternative hypothesis, specifying a real difference between the two groups. Given two tests that maintain the correct 5% size under the null hypothesis, we would prefer the one that correctly rejects the null hypothesis for a higher proportion of datasets generated under the alternative hypothesis, since it has higher power.

Simulation studies comparing some of the methods we consider have been performed elsewhere. For example, Feng et al. (1996) compared LMM and GEE, as well as two other estimation methods. Datta and Satten (2005) compared their test with both the standard Wilcoxon test ignoring clustering and the Wilcoxon test on cluster means. Rosner et al. (2006b) compared their test with LMM and the Wilcoxon signed rank test. Larocque et al. (2010) compared their proposals with the Datta and Satten (2005) and Rosner et al. (2006b) tests. However, to provide a comprehensive view of the methods available to analyze clustered data, we wanted to compare methods from all of the categories discussed in Part 1.4, approaches A–D, on a common set of simulated datasets. We are unaware of any previous papers that perform this full range of comparisons.

Part 2.1: Generation of simulated data

In preparing our simulated data, the number of clusters and the total number of observations were fixed for each simulation. Cluster sizes were generated from a multinomial distribution. We considered separately the two types of data distinguished in Part 1.2: case 1, where only a single group is represented in each cluster, and case 2, where clusters contain observations from both groups. For case 1, half of the clusters were assigned to each of the two groups of observations being compared. For case 2, approximately half the observations in a cluster were assigned to each of the two groups (exactly half, if the cluster size was even, or an imbalance of 1, if the cluster size was odd). Cluster sizes and type allocations were then held fixed for each simulation.

Normally distributed data were generated according to the model $y_{ik} = \mu + \beta z_{ik} + b_k + \varepsilon_{ik}$, and skewed data were generated according to the model $y_{ik} = \exp(\mu + b_k + \varepsilon_{ik}) + \beta z_{ik}$, where y_{ik} is the value of the response variable for unit i in cluster k, z_{ik} represents the group to which observation i in cluster k belongs (0 for group 1, 1 for group 2), b_k are independent and normally distributed with mean zero and variance $\sigma_b^2 = 0.7$, and ε_{ik} are independent and normally distributed with mean zero and variance $\sigma_w^2 = 0.3$. This corresponds to an intracluster correlation of 0.7. In these models, the parameter β corresponds to the group effect, such that $\beta = 0$ corresponds to no difference between the two groups of observations. Table 1 shows the remaining parameter choices for each simulation. By

altering these parameters, we generated datasets from eight different models, which were examined using various statistical methods (see below).

Part 2.2: Methods of statistical analysis

Having generated clustered data under eight different models (Table 1), we will next demonstrate how effectively seven different established statistical methods performed in the analysis of these different simulated datasets. The statistical tests we examined are as follows:

(1) A 2-sample *t* test, applied to the two groups of individual observations. In this test, clustering is not accounted for, since each observation is treated as independent.
(2) A Wilcoxon rank-sum test, applied to the two groups of individual observations. Clustering is not accounted for, as above.
(3) A *t* test, applied to the cluster means. For case 1 data, each cluster was reduced to a single mean and a two-sample *t* test was performed. For case 2 data, each cluster was reduced to a pair of means (one for each group) and a paired *t* test was performed. In these tests, clustering is effectively removed by reducing the data to the cluster means.
(4) A Wilcoxon rank-sum test, applied to the cluster means (for case 1 data) or a Wilcoxon signed rank test, applied to the paired cluster means (for case 2 data). Clustering is effectively removed by reducing the data to the cluster means, as above.
(5) LMM (Laird and Ware, 1982). This modeling approach incorporates the intracluster correlation effect, and thus accounts for data clustering without reducing the data to the cluster means.
(6) GEE (Liang and Zeger, 1986), using an "exchangeable" working correlation structure, i.e., correctly assuming equal correlation between observations within a cluster. This modeling approach also accounts for data clustering.
(7) Rank-sum test for clustered data (Datta and Satten, 2005). This is a modification of the Wilcoxon rank-sum test that accounts for the intracluster correlation.

LMM and *t* tests assume that the data are approximately normally distributed, whereas the other approaches do not require this assumption.

For case 2 data, we also considered a normal linear regression model (i.e., ANOVA, as described in Part 1.4, approach C) with a factor for cluster as well as the group effect. This is similar to LMM, except that cluster is treated as a fixed effect rather than a random effect. Since the results were very similar to LMM for the scenarios we considered, where clusters were approximately equally split between the two groups, we have not reported them separately. Note that this approach cannot be used for case 1 data, since there is no within-cluster comparison of the groups available and, hence, insufficient information to estimate both the group effect and a fixed effect for each cluster.

Part 2.3: Outcome of comparing statistical tests

A single simulated dataset is shown in Figure 1. This dataset consists of 200 observations in 20 clusters. The observations fall into two categories, group 1 and group 2, with only a single group represented in each cluster (i.e., case 1 data). In this simulated dataset, we have specified that there is no difference between group 1 and group 2 observations.

We compared group 1 against group 2 observations using the methods described above (Table 2). We observed that ignoring the clustering (applying the *t* test or Wilcoxon test to the individual observations) would lead to rejection of the null hypothesis, since the *p* values are considerably <0.05. Thus, when clustering was ignored, both the *t* and Wilcoxon tests reported a significant difference between group 1 and group 2 data, where in fact no such difference actually exists. None of the other methods indicated that the null hypothesis should be rejected. Thus, although the estimated difference between groups was similar for all of the methods,

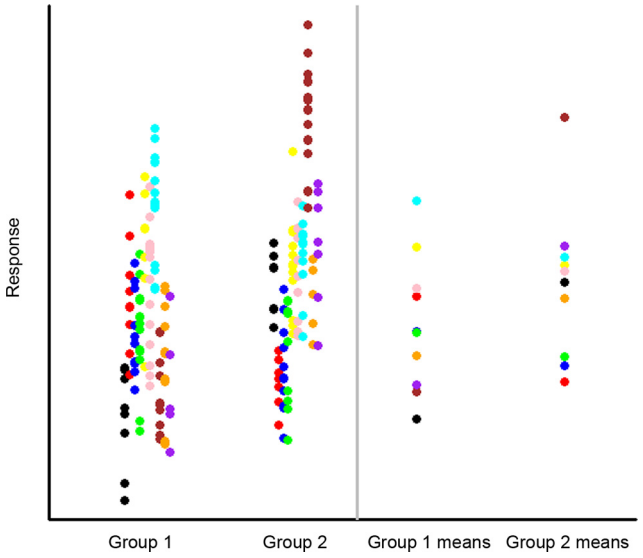

Figure 1. A sample dataset generated from model 1. For illustrative purposes, a single dataset generated under model 1 is shown. On the left are shown the individual observations using different colors for the different clusters within each group. In this simulation, no difference exists between group 1 and group 2 data. However, the similarity of observations within a cluster, which induces intracluster correlation, is apparent here, as observations within each cluster lie close to each other. On the right are shown the data reduced to cluster means. The color of each mean represents the cluster to which it belongs.

Table 2. Performance of various tests in analyzing sample data

Method	Estimated difference	95% CI for difference	*p* value for test of no difference
1. Two-sample *t* test (individual observations)	0.517	(0.230, 0.805)	0.0005
2. Wilcoxon (individual observations)	0.486	(0.200, 0.769)	0.0013
3. Two-sample *t* test (means)	0.476	(−0.318, 1.269)	0.2235
4. Wilcoxon (means)	0.425	(−0.387, 1.286)	0.2475
5. LMM	0.476	(−0.322, 1.275)	0.2260
6. GEE	0.476	(−0.228, 1.180)	0.1850
7. Rank-sum test for clustered data (Datta and Satten)			0.2015

Table 2 summarizes the results of applying the above methods to the dataset pictured in Fig. 1. We tested the null hypothesis that there was no significant difference between group 1 and group 2 observations. Methods 1 (*t* test) and 2 (Wilcoxon) were conducted on individual observations, thereby ignoring clustering. Methods 3 and 4 were applied after the data were reduced to the means of each cluster, thereby eliminating clustering. Unlike the *t* and Wilcoxon tests, LMM, GEE, and the rank-sum test of Datta and Satten (2005) are able to explicitly account for clustering. CI, Confidence interval.

Table 3. Performance of tests in analyzing 10,000 datasets, case 1 data

	Proportion of datasets for which null hypothesis is rejected						
Model	*t* test (individual observations)	Wilcoxon (individual observations)	*t* test (means)	Wilcoxon (means)	LMM	GEE	Datta and Satten method
1	0.512	0.496	0.047	0.043	0.049	0.079	0.053
2	0.966	0.961	0.688	0.641	0.692	0.772	0.699
3	0.500	0.496	0.031	0.042	0.037	0.067	0.053
4	0.781	0.960	0.359	0.529	0.367	0.448	0.706
5	0.431	0.412	0.041	0.030	0.050	0.144	0.065
6	0.986	0.981	0.723	0.604	0.770	0.927	0.806
7	0.405	0.412	0.021	0.029	0.034	0.124	0.065
8	0.892	0.960	0.569	0.568	0.625	0.786	0.735

Table summarizes the results (proportion of datasets for which the null hypothesis is rejected) when clusters contain observations from a single group only; data were obtained from 10,000 simulated datasets for each of the eight models described in Table 1 and for each of the analysis methods in Table 2. A significance level of 5% (i.e. $p < 0.05$) was used to determine whether to reject the null hypothesis in all cases.

Table 4. Performance of tests in analyzing 10000 datasets, case 2 data

	Proportion of datasets for which null hypothesis is rejected						
Model	*t* test (individual observations)	Wilcoxon (individual observations)	*t* test (means)	Wilcoxon (means)	LMM	GEE	Datta and Satten method
1	0.001	0.002	0.052	0.049	0.051	0.079	0.042
2	0.202	0.197	0.626	0.621	0.718	0.738	0.469
3	0.005	0.002	0.039	0.051	0.047	0.062	0.042
4	0.191	0.783	0.438	0.647	0.446	0.526	0.877
5	0.005	0.006	0.048	0.039	0.050	0.116	0.027
6	0.420	0.405	0.709	0.645	0.868	0.895	0.363
7	0.007	0.006	0.028	0.040	0.038	0.091	0.027
8	0.603	0.919	0.730	0.722	0.796	0.843	0.800

Table summarizes the results (proportion of datasets for which the null hypothesis is rejected) when clusters contain observations from both groups; data were obtained for 10000 simulated datasets for each of the eight models described in Table 1 and for each of the analysis methods in Table 2. A significance level of 5% (i.e. $p < 0.05$) was used to determine whether to reject the null hypothesis in all cases.

the *t* and Wilcoxon tests in which clustering was ignored underestimated the variability in the data due to the positive correlation in the data, as shown by the narrower confidence intervals. These findings demonstrate the hazardous nature of ignoring data clustering, since the tests in which clustering was ignored would report a significant difference between groups 1 and 2 when no such difference actually exists.

We next generated 10,000 datasets from each model detailed in Table 1. Analyzing such a large number of datasets provides a more comprehensive view of how the tests perform. Results for case 1 data, where only a single group is represented in each cluster, are shown in Table 3, while Table 4 shows results for case 2, in which clusters contain approximately equal numbers of observations from each group. For data generated under models 1, 3, 5, and 7 there is no real difference between the groups being compared, so that a valid statistical test should reject the null hypothesis for ~5% of the datasets.

Table 3 shows that for case 1 data, the *t* and Wilcoxon tests on individual observations (i.e., ignoring clustering) were far too liberal, rejecting the null hypothesis for >40% of the datasets under each null model. These two tests exhibited high power when there was a significant difference between the groups (models 2, 4, 6, and 8, in which $\beta \neq 0$). However, this is no consolation, because for a real dataset we would not know whether the null hypothesis is true or not. Thus, the probability of recording a "false positive" result would be much higher in tests where clustering is simply ignored. By contrast, Table 4 shows that for case 2 data, the *t* and Wilcoxon tests on individual observations were too conservative, rejecting the null hypothesis <1% of the time. These results are consistent with the mathematical findings presented in the supplemental Fig. S1, available at www.jneurosci.org as supplemental material. As the strength of the intracluster correlation increases, the effects of ignoring clustering become more pronounced. This is clear when the type I error is plotted against the intracluster correlation for case 1 and case 2 data (supplemental Fig. S1, available at www.jneurosci.org as supplemental material).

We then considered the performance of statistical methods that do not ignore clustering: *t* and Wilcoxon tests on cluster means, LMM, GEE, and the Datta and Satten (2005) method (described below).

Part 2.4: Large datasets, normally distributed data

Datasets from models 1–4 consist of 20 clusters and a total of 200 observations. Models 1 and 2 produce normally distributed data, and we expect that methods designed for normal data should perform well here. Table 3 shows that for case 1 data the LMM and *t* test on means, which are designed for normally distributed data, perform well. The Datta and Satten method, although not specifically designed for normal data, also performed well. These three methods controlled the type I error at close to 5% and achieved power of ~0.7. The Wilcoxon test on means had an acceptable type I error but was slightly less powerful.

Table 4 shows that for normally distributed case 2 data, in which clusters contain observations from both groups, LMM performed best, with type I error of 5% and power >0.7. The paired *t* and Wilcoxon tests performed similarly in terms of type I error but had slightly lower power. The Datta and Satten method had lower power in this case.

For both case 1 and case 2 data, GEE was slightly liberal when analyzing normally distributed data, with the type I error rate around 8% (Tables 3 and 4). The finding that hypothesis tests based on GEE tend to be too liberal when the number of clusters is relatively small has been noted previously (Feng et al., 1996; Mancl and DeRouen, 2001).

Part 2.5: Large datasets, non-normal data

For non-normal (skewed) data generated by models 3 and 4, the Datta and Satten method performed best, achieving the highest power while maintaining the correct rate of type I error (Tables 3 and 4).

Part 2.6: Smaller datasets

Datasets from models 5–8 are smaller, with 8 clusters and a total of 48 observations. For case 1 data, LMM performed the best for all model datasets, even for the non-normal datasets considered (Table 3). While the method of Datta and Satten also performed well in terms of power, it was somewhat more liberal with these smaller datasets than it was with the larger datasets of models 1–4.

For case 2 data, LMM clearly outperformed the other methods when analyzing normally distributed data (Table 4). It also performed well for skewed data, with a slightly conservative type I error but good power. The Datta and Satten method was also slightly conservative in this case but had relatively high power. The paired analyses on cluster means maintained type I error

<0.05 but had slightly lower power. GEE was even more liberal for the smaller datasets than for the larger datasets analyzed under Models 1–4.

Part 3: Examining the intracluster correlation

The strength of the intracluster correlation determines how similar observations within a given cluster are likely to be to each other. Thus, a higher intracluster correlation gives a more pronounced "clustering effect."

If the level of correlation itself is of interest, methods for estimating it are available. There are also methods for testing whether the correlation is significant, although even a small correlation, which a test may show to be nonsignificant, can have important implications for statistical analysis. We describe these methods in more detail in supplemental Fig. S2, available at www.jneurosci.org as supplemental material.

Part 4: Conclusions—which method to choose?

Clustered data are common in neuroscience research. Most studies have adopted one of two approaches to clustered data. The first of these is to ignore clustering entirely—an approach that can have profound implications for the outcomes of the analysis and can, as we have shown, sometimes lead to a conclusion that is incorrect. The consequences of failing to take clustering into account can be serious, even for a small amount of intracluster correlation. The second approach is reducing the data to a single measurement per cluster, usually by taking the mean of each cluster. While entirely valid, taking the cluster mean also reduces the amount of information that can be acquired from the dataset, as the data from the individual measurements in a cluster are reduced to one measurement. For the same reason, it may also be less powerful. We describe a number of alternative approaches to the analysis of clustered data and show that in many cases it is best to use a statistical method that explicitly accounts for data clustering.

As with more familiar methods of statistical analysis, the specific tests used depend very much on the nature of the data. We note therefore that our conclusions are limited to the models we investigated and may vary for other datasets. Our simulation studies, which are limited to continuous data, suggest that for large datasets LMM performs best for normally distributed data, while the rank-sum method of Datta and Satten (2005) performs well for non-normal data. Fixed linear regression/ANOVA can also be used as long as the data are comprised of clusters that contain observations from both of the groups being compared. LMM performed well for the smaller datasets that we analyzed; however, LMM may not do well for analyzing all types of skewed data. In addition, we specifically recommend, as with all statistical analyses, that larger datasets should always be studied where possible, since with very small datasets there may not exist a statistical method powerful enough to detect the difference of interest.

Software

Results were obtained using the R package (R Development Core Team, 2009). The standard two-sample *t* test and Wilcoxon tests are obtained from the t.test and wilcox.test functions in the stats package. LMM was implemented using the lme function in package nlme and GEE using the gee function in the gee package. The R code to implement the rank-sum test for clustered data developed by Datta and Satten (2005) is available from the authors.

Software for implementing some of the methods is also available in other packages, for example the SAS procedures MIXED for LMM and GENMOD for GEE.

References

Brown H, Prescott R (1999) Applied mixed models in medicine. Chichester, UK: Wiley.

Daniel JA, Galbraith S, Iacovitti L, Abdipranoto A, Vissel B (2009) Functional heterogeneity at dopamine release sites. J Neurosci 29:14670–14680.

Datta S, Satten GA (2005) Rank-sum tests for clustered data. J Am Stat Assoc 100:908–915.

Datta S, Satten GA (2008) A signed-rank test for clustered data. Biometrics 64:501–507.

Donner A, Banting D (1988) Analysis of site-specific data in dental studies. J Dent Res 67:1392–1395.

Feng ZD, McLerran D, Grizzle J (1996) A comparison of statistical methods for clustered data analysis with Gaussian error. Stat Med 15:1793–1806.

Fernandez-Alfonso T, Ryan TA (2004) The kinetics of synaptic vesicle pool depletion at CNS synaptic terminals. Neuron 41:943–953.

Gonen M, Panageas KS, Larson SM (2001) Statistical issues in analysis of diagnostic imaging experiments with multiple observations per patient. Radiology 221:763–767.

Krueger SR, Kolar A, Fitzsimonds RM (2003) The presynaptic release apparatus is functional in the absence of dendritic contact and highly mobile within isolated axons. Neuron 40:945–957.

Laird NM, Ware JH (1982) Random-effects models for longitudinal data. Biometrics 38:963–974.

Larocque D, Haataja R, Nevalainen J, Oja H (2010) Two sample tests for the nonparametric Behrens-Fisher problem with clustered data. J Nonparametric Stat. Advance online publication. Retrieved June 18, 2010. doi:10.1080/10485250903469728

Liang KY, Zeger SL (1986) Longitudinal data-analysis using generalized linear-models. Biometrika 73:13–22.

Mancl LA, DeRouen TA (2001) A covariance estimator for GEE with improved small-sample properties. Biometrics 57:126–134.

Moulder KL, Jiang X, Taylor AA, Shin W, Gillis KD, Mennerick S (2007) Vesicle pool heterogeneity at hippocampal glutamate and GABA synapses. J Neurosci 27:9846–9854.

Mozhayeva MG, Sara Y, Liu X, Kavalali ET (2002) Development of vesicle pools during maturation of hippocampal synapses. J Neurosci 22:654–665.

Murthy VN, Stevens CF (1998) Synaptic vesicles retain their identity through the endocytic cycle. Nature 392:497–501.

Murthy VN, Sejnowski TJ, Stevens CF (1997) Heterogeneous release properties of visualized individual hippocampal synapses. Neuron 18:599–612.

Pyle JL, Kavalali ET, Piedras-Renteria ES, Tsien RW (2000) Rapid reuse of readily releasable pool vesicles at hippocampal synapses. Neuron 28:221–231.

R Development Core Team (2009) R: A language and environment for statistical computing. Vienna: R Foundation for Statistical Computing.

Rosner B, Grove D (1999) Use of the Mann–Whitney U-test for clustered data. Stat Med 18:1387–1400.

Rosner B, Glynn RJ, Lee MLT (2003) Incorporation of clustering effects for the Wilcoxon rank sum test: A large-sample approach. Biometrics 59:1089–1098.

Rosner B, Glynn RJ, Lee MLT (2006a) The Wilcoxon signed rank test for paired comparisons of clustered data. Biometrics 62:185–192.

Rosner B, Glynn RJ, Lee MLT (2006b) Extension of the rank sum test for clustered data: two-group comparisons with group membership defined at the subunit level. Biometrics 62:1251–1259.

Ryan TA, Reuter H, Smith SJ (1997) Optical detection of a quantal presynaptic membrane turnover. Nature 388:478–482.

Senn S (1998) Some controversies in planning and analysing multicentre trials. Stat Med 17:1753–1765.

Virmani T, Atasoy D, Kavalali ET (2006) Synaptic vesicle recycling adapts to chronic changes in activity. J Neurosci 26:2197–2206.

Waters J, Smith SJ (2002) Vesicle pool partitioning influences presynaptic diversity and weighting in rat hippocampal synapses. J Physiol 541:811–823.

Willeumier K, Pulst SM, Schweizer FE (2006) Proteasome inhibition triggers activity-dependent increase in the size of the recycling vesicle pool in cultured hippocampal neurons. J Neurosci 26:11333–11341.

Zyzanski SJ, Flocke SA, Dickinson LM (2004) On the nature and analysis of clustered data. Ann Fam Med 2:199–200.