NeuroImage 277 (2023) 120232

Contents lists available at ScienceDirect

NeuroImage

journal homepage: www.elsevier.com/locate/neuroimage

Permutation-based group sequential analyses for cognitive neuroscience

John P. Veillette*, Letitia Ho, Howard C. Nusbaum

Department of Psychology, University of Chicago, United States

ARTICLE INFO

Keywords:
Inference
Sample size
Reproducibility

ABSTRACT

Cognitive neuroscientists have been grappling with two related experimental design problems. First, the complexity of neuroimaging data (e.g. often hundreds of thousands of correlated measurements) and analysis pipelines demands bespoke, non-parametric statistical tests for valid inference, and these tests often lack an agreed-upon method for performing a priori power analyses. Thus, sample size determination for neuroimaging studies is often arbitrary or inferred from other putatively but questionably similar studies, which can result in underpowered designs – undermining the efficacy of neuroimaging research. Second, when meta-analyses estimate the sample sizes required to obtain reasonable statistical power, estimated sample sizes can be prohibitively large given the resource constraints of many labs. We propose the use of sequential analyses to partially address both of these problems. Sequential study designs – in which the data is analyzed at interim points during data collection and data collection can be stopped if the planned test statistic satisfies a stopping rule specified a priori – are common in the clinical trial literature, due to the efficiency gains they afford over fixed-sample designs. However, the corrections used to control false positive rates in existing approaches to sequential testing rely on parametric assumptions that are often violated in neuroimaging settings. We introduce a general permutation scheme that allows sequential designs to be used with arbitrary test statistics. By simulation, we show that this scheme controls the false positive rate across multiple interim analyses. Then, performing power analyses for seven evoked response effects seen in the EEG literature, we show that this sequential analysis approach can substantially outperform fixed-sample approaches (i.e. require fewer subjects, on average, to detect a true effect) when study designs are sufficiently well-powered. To facilitate the adoption of this methodology, we provide a Python package "niseq" with sequential implementations of common tests used for neuroimaging: cluster-based permutation tests, threshold-free cluster enhancement, *t*-max, *F*-max, and the network-based statistic with tutorial examples using EEG and fMRI data.

1. Introduction

Recently, many scientific fields have been placing a renewed emphasis on issues of sample size determination and statistical power. This emphasis is motivated, in large part, by an increased appreciation of the fact that the positive predictive value of a study – that is, the probability that an effect is actually "true" given a statistically significant result – is directly proportional to the statistical power of the study (Ioannidis, 2005). This methodological concern came into focus in the neurosciences after a landmark review in 2013 estimated that the median statistical power of neuroscience studies is between 8% and 31%, suggesting that many neuroscience studies provide low evidentiary value despite satisfying conventional standards of statistical evidence (Button et al., 2013).

The average statistical power of neuroimaging studies, in particular, has been steadily improving; however, low statistical power is cited as one of the largest threats to the replicability of findings in cognitive neuroscience (Poldrack et al., 2017). The availability of large, open datasets such as the Human Connectome Project (Van Essen et al., 2013) and tools for extracting metadata from published neuroimaging studies such as NeuroSynth (Yarkoni et al., 2011) have enabled empirical estimates of power in the field. As a result, we now know that the effect sizes one can realistically expect in neuroimaging studies are usually quite small, and many published neuroimaging studies are too small to detect them (Poldrack et al., 2017). Indeed, a recent analysis has suggested that certain types of neuroimaging biomarkers may even require *thousands* of subjects to detect reliably (Marek et al., 2022), though other researchers have been quick to point out that not all analytic approaches require such prohibitive sample sizes to establish robust brain-behavior relationships (Rosenberg and Finn, 2022). In any event, the costs of doing well-powered neuroimaging research can be substantial.

In light of these field-wide concerns, it is increasingly acknowledged that researchers should determine their sample-size in a rigorous, non-arbitrary manner; the use of heuristics, such as adapting the same sam-

* Corresponding author.
E-mail address: johnv@uchicago.edu (J.P. Veillette).

https://doi.org/10.1016/j.neuroimage.2023.120232.
Received 15 May 2023; Received in revised form 13 June 2023; Accepted 15 June 2023
Available online 20 June 2023.

ple size as a previous study, is likely to result in an underpowered design (Poldrack et al., 2017). Some efforts – notably Neuropower, Fmripower, and PowerMap – have provided researchers with the tools to perform power analyses for the parametric random-field theory approaches used in fMRI (Durnez et al., 2016; Joyce and Hayasaka, 2012; Mumford and Nichols, 2008). However, these tools remain limited relative to the scope of statistical tests employed in the broader neuroimaging literature. In particular, many of the statistical tests used in neuroimaging – cluster-based permutation tests (Maris and Oostenveld, 2007), threshold-free cluster enhancement (Smith and Nichols, 2009), the network-based statistic (Zalesky et al., 2010), and *t*-max (Nichols and Holmes, 2002), to name a few – are non-parametric, and thus parametric power analysis procedures are inapplicable (though a power analysis may be performed by simulation or resampling). Moreover, even in cases in which a validated method for performing a power analysis would be straightforward, specifying an "effect size" is not as simple as for univariate tests, in which power analyses are performed with standardized effect size measures (e.g. Cohen's *d*); neuroimaging analyses are often performed on aggregate properties of spatiotemporal maps, not necessarily on specific voxels. Indeed, an effect map (which would be difficult to predict a priori) is often precisely what the researcher is trying to estimate when they are designing their experiment. Moreover, even in the univariate case, obtaining a reasonably precise effect size estimate usually requires a larger sample size than that required to merely detect an effect (Albers and Lakens, 2018; Lakens and Evers, 2014), and it understandably belies most researchers' intuition to collect a pilot sample larger than their confirmatory study.

Most published fMRI and EEG studies still do not include a sample size justification, likely due to the substantial challenges associated with performing a power analysis described above. For instance, out of 100 clinical fMRI studies randomly sampled from six leading journals, only a single study reported a sample size calculation (Guo et al., 2014). Similarly, 0 out of 100 randomly sampled studies from the EEG literature reported sample size calculations (Larson and Carbine, 2017). We do not believe that this omission results from researchers' lack of desire to do more rigorous science, but rather it is the result of a lack of methodological approaches and tools that meet their specific research needs. Indeed, when reviewers for granting agencies such as NSF and NIH request sample size calculations for proposed research, PIs may scramble to find a way of estimating these, whether rigorous or not.

One alternative researchers may find tempting is to forgo an a priori sample size determination and analyze data multiple times throughout the course of data collection, stopping data collection only once a significant result is found. Indeed, many researchers in psychology admit to employing this practice known as *optional stopping* (John et al., 2012), one of a laundry list of "questionable research practices" such as "*p*-hacking" that have come under criticism in recent years (de Vrieze, 2021). However, optional stopping results in inflated false-positive rates; for instance, if one analyzes the data five times throughout data collection without adjusting their significance threshold of $\alpha = 0.05$, the false positive rate rises to an undesired 0.142 (Armitage et al., 1969). Again, however, we do not believe researchers admit to optional stopping because of a negligent disregard for research best-practices, but because of a desire to preserve time and resources, stopping data collection as soon as there is sufficient evidence to reach a conclusion. Indeed, one could argue researchers have an ethical obligation to use (often taxpayer-funded) resources efficiently and to limit the burden on the human subject populations that volunteer for their experiments.

Fortunately, valid *sequential analysis* techniques, which use adjusted rejection thresholds at each interim analysis to control the false positive rate across the whole experiment, have existed for the better part of a century (Dodge and Romig, 1929; Wald, 1992). Sequential study designs have long been recognized to possess efficiency advantages over fixed-sample designs, allowing a conclusion to be reached with fewer observations on average (Dodge and Romig, 1929). The reason for this efficiency advantage is straightforward. For instance, if one analyzes the data once at a sample size n_1 where there is a 50% chance of rejecting the null hypothesis and once at n_2 when there is a 95% chance of rejection, then the expected sample size at which a conclusion is reached would likely end up somewhere between n_1 and n_2, since the probability of rejection at n_1 is substantial.

While early approaches to sequential analysis required data to be analyzed after every observation or, alternatively, the exact number and timing of looks at the data to be specified a priori (a *group* sequential analysis), which was somewhat limiting for studies where the final subject yield may not be known until the data are analyzed (e.g. after quality check of neuroimaging data), a later approach known as *alpha spending* requires only that a maximum sample size (that is, the sample size at which data collection will be terminated even if a significant result has not been reached) be specified ahead of time (Lan and DeMets, 1983). This approach is heavily used in the clinical trial literature and has recently attracted attention as a means of improving the efficiency of experimental psychology studies as well (Lakens, 2014). While we believe the general alpha spending approach is sufficiently flexible to meet the practical demands of neuroimaging studies, the adjusted significance thresholds it prescribes for interim analyses are only valid assuming test statistics across looks follow a multivariate normal distribution (Lan and DeMets, 1983), an assumption that is violated by many of the test statistics in neuroimaging.

To this end, we propose a permutation-based version of the alpha spending procedure. This procedure allows sequential analyses to be performed using arbitrary test statistics, and we show by simulation that it controls the false positive rate while doing so. Since many hypothesis tests and multiple-comparisons corrections used in cognitive neuroscience are already permutation-based, they can be naturally generalized into valid sequential tests. Thus, we were able to implement sequential generalizations of cluster-based permutation tests (Maris and Oostenveld, 2007), threshold-free cluster enhancement (Smith and Nichols, 2009), *t*-max and *F*-max corrections (Nichols and Holmes, 2002), and the network-based statistic (Zalesky et al., 2010). As such, the proposed approach to sequential-testing can be easily applied to task EEG and fMRI, functional connectivity, diffusion tractography, and voxel-based morphometry data.

This sequential testing approach affords several routes to principled sample size determination that were previously unavailable to human neuroscience researchers. (1) One approach is to specify a very conservative maximum sample size; if the maximum sample is overly conservative, one will reject the null hypothesis at an interim analysis with high probability. (2) Another approach is to use an *adaptive design*, in which the maximum sample size can be adjusted as a result of a *conditional power analysis* performed at an interim analysis (Lakens et al., 2021). In this case, the first interim analysis can serve as an internal pilot study, used to estimate the effect size with which a power analysis is performed (which can be done very generally via resampling methods). However, if enough evidence has already been accrued at that interim analysis to reject the null hypothesis, then data collection can be concluded. (3) Lastly, even if a researcher has everything needed to perform an a priori sample size calculation, it may still be advantageous to use a sequential design so that data collection can be terminated early if enough evidence has been accrued to reject the null hypothesis. We show that, like parametric alpha spending, our approach can be substantially more efficient than fixed-sample designs for studies that are sufficiently well-powered.

2. Methods

2.1. Alpha spending

2.1.1. General description

In an alpha spending procedure, one must decide two things in advance of beginning data collection. (1) One must specify a maximum sample size after which data collection will be stopped, regardless of

whether the null hypothesis has been rejected. We will call this design parameter n_T throughout this article, where T denotes the total number of interim analyses. Note, however, that T does not need to be determined before the study begins, only $n_T = n_{\max}$. (2) Additionally, one must define a desired Type I error rate α and an *alpha spending function* that specifies how the Type I error rate will be distributed across interim analyses. Specifically, an alpha spending function $s(n)$ is a monotonically non-decreasing function that specifies a target value for the *cumulative* Type I error by interim sample size n. Consequently, the function s must also satisfy $s(0) = 0$ and $s(n_T) = \alpha$ for the desired significance level α, such that the Type I error rate is contained to α at n_T. Within these constraints, any function $s: \mathbb{R}_1 \to \mathbb{R}_1$ (i.e. that maps a real number to a real number) could be specified.

Suppose the data are analyzed at each of the successive interim sample sizes $\{n_t\}_{t=1}^{T}$. (This notation more concisely denotes a set containing ordered interim sample sizes $n_1, n_2, \ldots, n_{T-1}, n_T$ for an arbitrary positive integer T.) At each interim sample size n_i, then, the data are $X_t = \{x_j\}_{j=1}^{n_t}$ (where the index of x_i, importantly, denotes the order in which the observation was collected rather than being a random index) with condition labels (or covariate of interest) $Y_t = \{y_j\}_{j=1}^{n_t}$, and the test statistic is $F_t = f(X_t, Y_t)$, where f is some function that computes the test statistic from the data.

We need to determine some threshold F_i^* at each interim analysis that will contain the cumulative Type I error, given all the previous looks, to $s(n_t)$. That is, for each interim analysis $t \in [1, T]$, we select F_t^* so as to satisfy Equation 1 under the null hypothesis.

$$s(n_t) = P\left(\bigcup_{t}^{j=1} F_j \geq F_j^*\right) \tag{1}$$

Note that the right side of Equation 1 can be expanded as in Equation 2.

$$P\left(\bigcup_{t}^{j=1} F_j \geq F_j^*\right) = P\left(F_t \geq F_t^*\right) + P\left(\bigcup_{t-1}^{j=1} F_j \geq F_j^*\right) - P\left(F_t \geq F_t^* \cap \bigcup_{t-1}^{j=1} F_j \geq F_j^*\right) \tag{2}$$

It is this joint probability on the right side of Equation 2 that can be difficult to compute, but it will almost certainly be nonzero since the F_t's are computed from overlapping data. In the parametric approach to alpha spending introduced by Lan and DeMets, the joint distribution of the test statistics $\{F_t\}_{t=1}^{T}$ is assumed to be multivariate normal under the null hypothesis, which allows appropriate rejection thresholds $\{F_t^*\}_{t=1}^{T}$ to be computed exactly. In the event that this normality assumption is violated, these thresholds must be estimated in some other way.

Note that, by plugging the rejection thresholds $\{F_t^*\}_{t=1}^{T}$ into the cumulative distribution function of F_t under the null hypothesis, it is possible to obtain adjusted significance levels $\{\alpha_t\}_{t=1}^{T}$ for each look time. It is common to provide these adjusted significance levels when reporting the results of a sequential analysis (Lakens et al., 2021). An example of adjusted signifance levels for a sequential design, as determined by an alpha spending procedure, is visualized in Fig. 1. pvalues

2.1.2. Common spending functions

As noted above, there is considerable flexibility in the selection of an alpha spending function $s(n)$; any function that satisfies the constraints specified in the above section will be adequate. However, it is worth mentioning a few common spending functions used in the literature, visualized in Fig. 2. Lan and DeMets introduced two spending functions, the Pocock and O'Brien Fleming spending functions (Lan and DeMets, 1983). The Pocock spending function "spends" the Type I error rate more liberally early during data collection, so it is more likely to reject the null hypothesis early on if the effect size is large. The O'Brien Fleming spending function, on the other hand, "saves" its Type I error allotment for later during data collection when there is more power to detect a given effect, so while it is less likely to terminate very early if the effect size is unexpectedly large, it may be more likely to reject the null hypothesis at an interim analysis overall. A linear spending function, which falls somewhere between those two, distributed the error rate evenly across data collection. We recommend consulting Lakens and colleagues' tutorial on group sequential designs for a more thorough discussion and comparison of adjusted alpha thresholds between spending functions (Lakens et al., 2021).

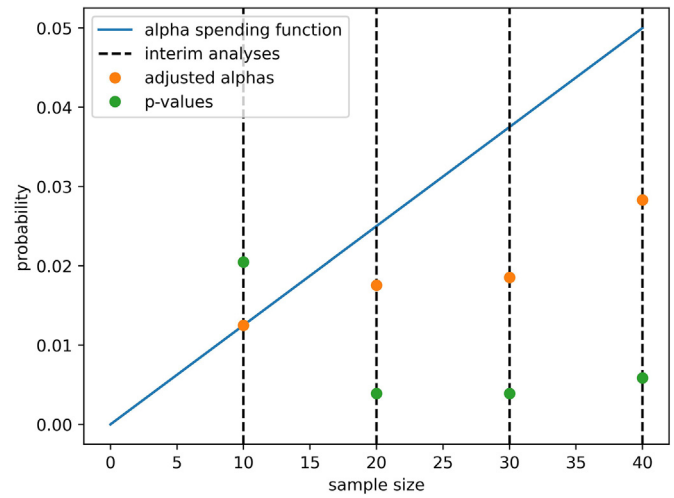

Fig. 1. An example of an alpha spending procedure with a linear spending function. At each interim analysis, an adjusted significance level α is computed that controls the *cumulative* Type I error, given the previous interim analyses, to the value specified by the alpha spending function. In this example, data collection would be stopped at $n = 20$, where the observed p-value drops below the adjusted α.

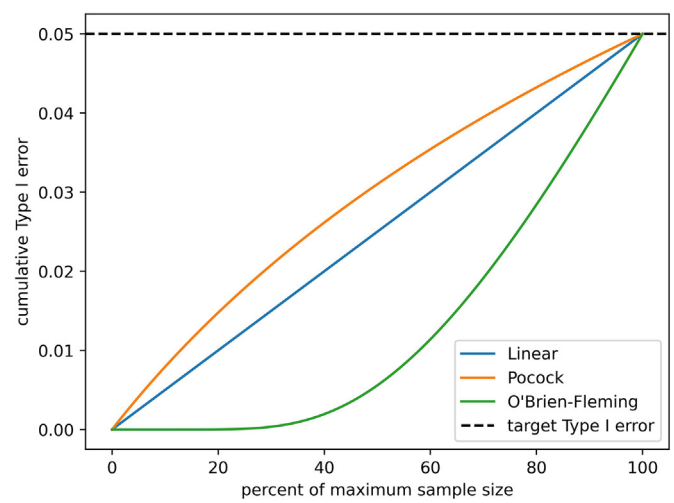

Fig. 2. Examples of common alpha spending functions. Spending functions all start at 0 and end at the user-specified false positive (Type I error) rate, but they vary in how they distribute Type I error across interim analyses.

2.1.3. Inflation factors and expected sample size

Generally speaking, if one wants to design a study using alpha spending that has the same statistical power as a given fixed sample design

with sample size n_{fixed}, then the maximum sample size n_T for the sequential design will need to be greater than n_{fixed}. The ratio n_T/n_{fixed} required for the two designs to have matched power is called the *inflation factor,* and it depends on the alpha spending function, the number and timing of interim analyses, and the desired statistical power. However, usefully, it does not depend on n_{fixed}, the effect size, or the statistical test used.

Even more usefully, the inflation factors for parametric alpha spending designs are known exactly, and they can be easily obtained from open-source statistical software such as rpact (Wassmer and Pahlke, 2020). While these inflation factors are not theoretically guaranteed to apply to the permutation alpha spending approach we describe below, empirically they seem to work quite well (see our simulation results below).

While the inflation factor n_T/n_{fixed} needed to obtain matched statistical power may generally be greater than one (i.e. the maximum sample size for the sequential design must be larger than the fixed-sample design), this inflation is offset by the fact that sequential designs have the opportunity to reject the null hypothesis at an interim analysis and therefore n_T observations may never actually be collected. The *expected sample size* of a sequential design, then, differs from the maximum sample size.

The expected sample size depends on the power of the sequential design. If the study is sufficiently well powered, the probability of detecting the effect early is high and the expected sample size is often substantially lower than the corresponding fixed-sample design's sample size; if, conversely, the design is very poorly powered, then the experiment will likely be run to completion without detecting the effect and the expected sample size will be close to n_T. The expected sample size of a given sequential design, as a function of statistical power, can be found exactly for parametric alpha spending designs and can be easily obtained from software packages like rpact.

2.2. Permutation alpha spending

We propose a permutation-based procedure to infer the co-dependence term in Equation 2 (the intersection probability following the minus sign) implicitly from the data. Conceptually, this is similar to how permutation-based approaches for controlling the familywise error rate outperform parametric approaches (e.g. Bonferroni correction) in the event that one's multiple tests are correlated, since the covariance between test statistics is implicitly accounted for because that correlation structure is preserved in the empirical null distribution generated by permutating the data (Nichols and Holmes, 2002). That is, the permutation null is a *joint* null distribution.

Let H_0 be a $K \times T$ matrix, where K is the number of permutations we decide to perform. We want to populate the matrix H_0 with the *joint null distribution* of the test statistic across the T interim analyses, which we will empirically estimate by permutation.

Then for each permutation $k \in [1,\ K]$ and each $t \in [1,\ T]$,

1. Let $Y_t^* = \{Y_{t-1}^* \cup \text{shuffled}(\{y_j\}_{j=n_{(t-1)}+1}^{n_t})\}$, where Y_0^* is an empty set.
2. Set $H_0^{(k,\ t)} = f(X_t,\ Y_t^*)$

Note that Step 1 above implies that any observation x_i has the same shuffled label at all times $t \in [1,\ T]$ on a given permutation k. Without this feature, H_0 will not be a joint distribution. If one is performing a one-sample test instead of an independent-samples tests, then the signs of the x_i's are randomly flipped instead of shuffling their y_i labels. (And similarly, a paired-sample test is just a one-sample test on the paired differences, so the signs of the differences are flipped, which is equivalent to randomly shuffling the labels within the pairs.)

One can then estimate the rejection thresholds from H_0 as follows.

- Pick F_1^* such that only $100 \cdot s(n_1)$ percent of the values in $H_0^{(:,\ 1)}$ (i.e. the first column of H_0) surpass F_1^*.
- In ascending order of t, pick each F_t^* such that only $100 \cdot s(n_t)$ percent of permutations k have any $H_0^{(k,\ j)}$ that exceed F_j^* at any $j \in [1, t]$.

That is, each F_t^* is chosen to control the *cumulative* false positive rate given all previous F^* to match the target given by the alpha spending function $s(n_t)$. Since all thresholds F_t^* depend on the previous thresholds and data, but not on the later thresholds and data, then rejection thresholds can be computed at the time of each interim analyses t, and data collection can be halted if $F_t \geq F_t^*$.

As in the parametric case, adjusted α_t's can be obtained from each F_t^* by computing the percentile rank of F_t^* within $H_0^{(:,\ t)}$, which is approximately equivalent to plugging F_t^* into the empirical cumulative distribution function of F_t under the null hypothesis.

2.3. Simulation studies

2.3.1. False positive rate simulations

First, we validate the permutation scheme by estimating the false positive rate for a univariate test by Monte Carlo simulation. In particular, on each simulation, we generate 500 random variables from the standard Normal distribution. We then compare two approaches. (1) In a simulated optional stopping procedure, we compute a p-value using a two-sided, one-sample permutation t-test with 1024 permutations at each of $n = 100,\ 200,\ 300,\ 400,\ 500$, and we reject the null hypotheses if $p \leq 0.05$ at any of these interim analyses. Estimating the false positive rate as the proportion of 10,000 simulations in which the null hypothesis was erroneously rejected, the false positive rate of this optional stopping procedure should be about 0.142 (Armitage et al., 1969). (2) We perform the same interim analyses, but we compute adjusted significance thresholds using the permutation alpha spending procedure described above with a linear spending function. Then, we only reject the null hypothesis if a test statistic exceeds the adjusted threshold at an interim analysis. This procedure should result in a false positive rate of approximately $\alpha = 0.05$.

We additionally perform the above simulations for a between-sample test, in which the simulated observations are randomly assigned to one of two conditions and an independent samples permutation t-test is used to compare between conditions.

We also verify that this procedure still controls the false positive rate for more sophisticated test statistics (e.g. one actually used on effect maps in neuroimaging). This time, we repeat all the above simulations but instead of generating random observations of one variable, we generate observations from 100 variables (i.e. "voxels") resulting in 100 different tests at each interim analysis. We control for multiple comparisons at each look time using a t-max procedure (Nichols and Holmes, 2002). (1) In the optional stopping procedure, we reject the null hypothesis if a t-max adjusted p-value (from any of the 100 tests) is less than or equal to 0.05 at any interim analysis. (2) In the sequential analysis procedure, we compute adjusted significance levels using permutation alpha spending with the t-max statistic, and we only reject the null hypothesis if a t-max adjusted p-value is less than or equal to those adjusted thresholds.

Finally, since the simulations described above use standard normal random variates, but we are also interested in cases where the data may be non-normally distributed, we also perform paired- and independent-sample t-max simulations as described above using chi-squared random variables with five degrees of freedom. Since the values in a power spectrum computed with the Fourier transform theoretically follow a chi-squared distribution, this is actually quite a typical family of distributions to encounter in M/EEG data analysis.

Note that, since the Type I error rate of the t-max procedure does not depend on the correlation between voxel-level tests (Smith and Nichols, 2009), we use independent tests here for computational expediency. The Type II error rate (i.e. power), however, is sensitive to the correlation structure of the data, so we use real EEG data in the simulations that follow (see Section 2.3.2 below).

2.3.2. Power/efficiency simulations

To assess the efficiency of sequential designs relative to fixed-sample designs, we ran power analyses for detecting seven canonical EEG event-related potential (ERP) effects from the ERP CORE dataset (Kappenman et al., 2021) using both types of design. Specifically, we used the lowpass-filtered versions of the precomputed difference waves for each ERP effect provided in ERP CORE. The dataset contains data from 40 healthy subjects for each ERP effect, though some of the subjects are excluded from certain ERP effects for low quality data. See the ERP CORE reference paper for a full description of data collection and preprocessing (Kappenman et al., 2021).

To estimate the statistical power for a given sample size n, we use a resampling technique. Specifically, we use a modified Bayesian bootstrap, which differs from the frequentist bootstrap in that it simulates draws from the posterior distribution (with uninformative priors) of a parameter instead of the sampling distribution of that parameter (Rubin, 1981). Strictly speaking, the frequentist bootstrap is a special case of the Bayesian bootstrap in which the original observations are resampled with the same probability on every bootstrap resampling; the Bayesian bootstrap draws the new resampling probabilities from a Dirichlet distribution on each bootstrap resampling. The result is that "the Bayesian bootstrap can be thought of as a smoothed version of the Efron bootstrap" (Lancaster, 2003), generally yielding more stable results when resampling from smaller pools of observations – though it asymptotically converges with the frequentist bootstrap. Estimating power by Bayesian bootstrap lends itself to interpretation as a non-parametric Bayesian predictive power (Spiegelhalter et al., 1986), which appropriately accounts for uncertainty about the effect size in estimating the power of a frequentist test.

On each simulation (i.e. bootstrap resample), then, we resample n observations from the original 40 observations and run our statistical test. Power is estimated as the proportion of simulations in which the null hypothesis is rejected.

For each ERP effect, we estimate the statistical power of a fixed-sample design with $n = 30$. Then, we estimate the maximum sample size n_T needed to obtain the same power using a sequential design – with a single interim analysis performed midway through data collection and a Pocock spending function – by obtaining the necessary inflation factor (see Section 2.1.2) from rpact (Wassmer and Pahlke, 2020). We then estimate the statistical power of the sequential design in the same manner (i.e. the proportion of simulations in which the null hypothesis is rejected at either the interim or maximum sample size), and we additionally estimate the expected sample size (see Section 2.1.2) as the average of the sample size at which "data collection" was terminated (either because the null hypothesis was rejected or the simulated study was completed without rejecting the null) across all simulations.

Since the ERP CORE consists of highly optimized EEG paradigms, designed to maximally elicit the ERP effect of interest, it is sometimes the case that every subject in the dataset shows the effect individually. If this is the case, our bootstrap procedure might estimate a power of 1, which is not informative for our purposes since an inflation factor cannot be computed for a design of power 1. So, when this occurs, we modify the bootstrap procedure as follows. After drawing our n resamples from the original observations on each bootstrap simulation, we generate a n noise time series (see next paragraph for noise generation procedure) and add them to the n samples. If this addition of noise was not sufficient to decrease power to be less than 1, we ran the simulations again, multiplying the noise by 2. If this failed, we multiplied the noise by 3. No effect required noise be multiplied by a factor greater than 3. The exact noise multipliers used for each effect can be found in our archived results (see Section 2.4: *Data and Code Availability*).

To generate noise, we estimate the covariance (between sensors) matrix of the grand-averaged difference wave for the ERP effect; then, we generate spatially colored multivariate Gaussian white noise using this covariance matrix so as to maximally interfere with the effect of interest. We then lowpass filtered the noise to match the ERP data (which was filtered at 20 Hz), which created temporal autocorrelation in the filtered signal.

10,000 bootstrap simulations were performed as above for each ERP effect using both a t-max test and a cluster-based permutation test (with a clustering threshold of $t = 2$), both with 1024 permutations. After running all simulations, (a) we compared the power of the fixed sample designs to that of the sequential designs, which should have a 1:1 relationship if the inflation factors adequately predict the sample size required to match the power of a fixed sample design. (b) Then, we assess the efficiency of the sequential designs by comparing their expected sample sizes to the fixed sample size, which is always 30.

2.4. Data and code availability

Our implementations of permutation alpha spending for cluster-based permutation tests, threshold-free cluster enhancement, t-max, F-max, r-max, and the network-based statistic are contained in our user-friendly Python package *niseq*, which can be installed from the Python package index (PyPI). Documentation is hosted on Read the Docs (http://niseq.readthedocs.io/). Source code, as well as worked examples using the package on EEG and fMRI data in conjunction with the MNE-Python (Gramfort et al., 2014) and nilearn packages, are available on GitHub (https://github.com/john-veillette/niseq) and permanently archived on Zenodo. The most recent release as of writing (v0.0.2) is available at https://doi.org/10.5281/zenodo.7526535 and the current release is always archived at https://doi.org/10.5281/zenodo.7517285.

The code used for the simulations featured in this article, as well as the results of those simulations and a record of the simulation parameters, are available separately on GitHub (https://github.com/john-veillette/niseq-simulations) and are permanently archived on Zenodo (https://doi.org/10.5281/zenodo.7666443).

The data used for simulations was originally taken from the ERP CORE repository on the Open Science Framework (https://doi.org/10.18115/D5JW4R). However, we exported the preprocessed difference waves into a file format that could be easily loaded with MNE-Python, which we provide for convenience in the same Zenodo archive as our simulation code.

2.5. Ethics statement

The ERP CORE dataset used in our simulations was collected with approval from the Institutional Review Board at the University of California, Davis, and all participants provided informed consent (Kappenman et al., 2021). Similarly, the Brainomics dataset used in the tutorial example was collected with approval from a regional ethics committee (Hopital de Bicêtre, France), and all subjects provided informed consent (Papadopoulos Orfanos et al., 2017; Pinel et al., 2007).

3. Results

3.1. False positive rates

Permutation alpha spending controls the false positive rate below the specified $\alpha = 0.05$ across multiple interim analyses for single permutation t-tests, and it controls the familywise error rate in a sequential t-max procedure (see Table 1). In contrast, permutation t-tests and t-max with optional stopping results in inflated false positive rates.

Notably, how to best correct for multiple comparisons in group sequential designs with multiple outcomes of interest is still an active topic of research in the clinical trial literature (Glimm et al., 2010; Kosorok et al., 2004; Tang and Geller, 1999). Interestingly, even when controlling for multiple comparisons at each interim analysis, out simulations suggest false positive rate inflation due to optional stopping is worsened in a multiple testing setting. Sequential t-max provides a solution to this problem that can scale to hundreds and thousands of arbitrarily correlated tests.

Table 1
False positive rates for optional stopping and for permutation alpha spending procedures with five interim analyses. For *t*-max, the false positive rate reported is a familywise error rate (probability of at least one Type I error across 100 tests and five interim analyses).

	One sample (univariate)	One sample (*t*-max)	Paired (*t*-max, chi-square)	Independent (univariate)	Independent (*t*-max)	Independent (*t*-max, chi-square)
Optional stopping	0.1442	0.1784	0.1794	0.1402	0.1715	0.1812
Alpha spending	0.0470	0.0439	0.0452	0.0465	0.0458	0.0486

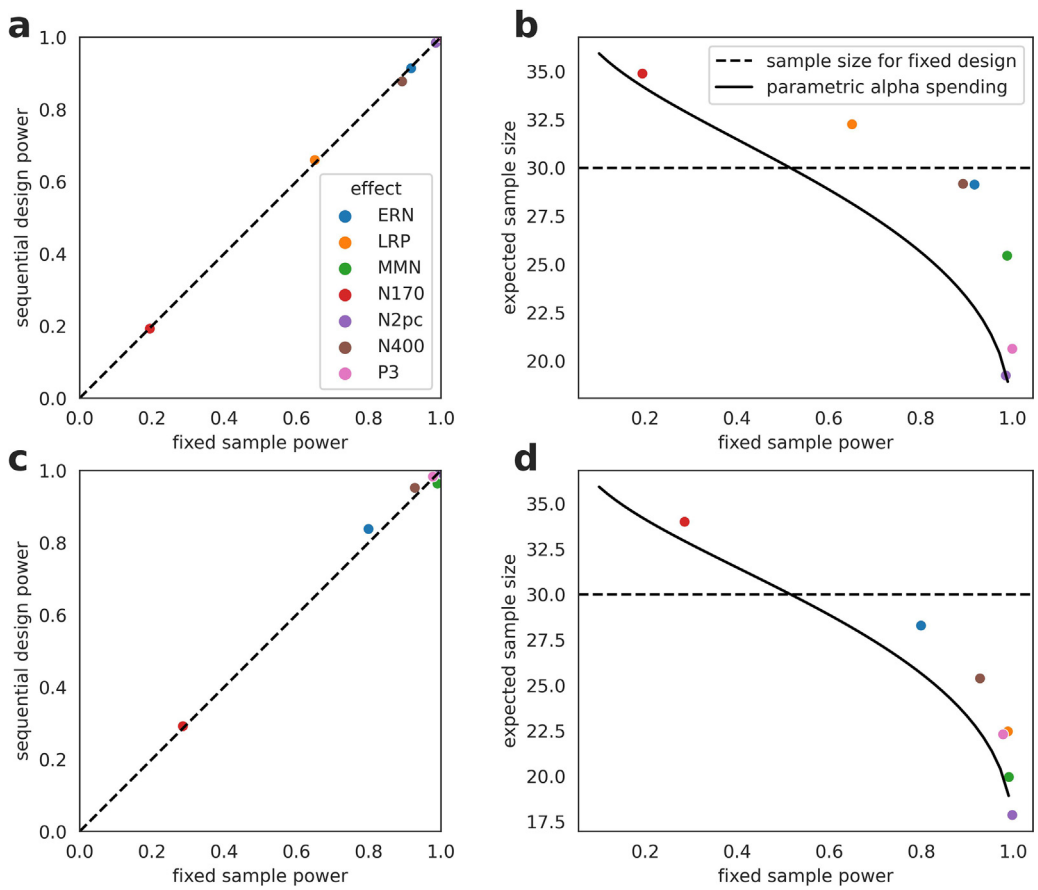

Fig. 3. Results of power simulations. (a and c) Power for detecting ERP effects compared between fixed-sample designs with $n_{\text{fixed}} = 30$ and sequential designs with one interim analysis, a Pocock spending function, and $n_{\max} = 30\times$ IF, where IF is an inflation factor known a priori (see Section 2.1.2). **(b and d)** The expected sample size of the sequential designs as a function of statistical power for detecting the effect of interest, compared to the corresponding sample-size vs. power curve for parametric sequential designs, which may be considered a lower limit for expected sample size. **(a) and (b)** show the results for a *t*-max test, while **(c) and (d)** show the results for a cluster-based permutation test. In all panels, each dot color corresponds to a different ERP effect in the ERP CORE dataset, see legend in **(a).**

3.2. Efficiency relative to fixed-sample designs

The results of our power simulations for fixed-sample and sequential designs are illustrated in Fig. 3. When we set the maximum sample size for each sequential design by multiplying the sample size of the fixed-sample design (always $n_{\text{fixed}} = 30$) by the appropriate inflation factor, we obtain a design with roughly matched power (see Fig. 3a and 3c). In other words, the inflation factors used for parametric sequential designs seem applicable to our approach.

When the study design is well-powered for detecting an effect, the expected sample size is smaller because the probability of rejecting the null hypothesis at an interim analysis is higher Consequently, for sufficiently well-powered designs, we see up to >30% sample size saving compared to a fixed-sample design with the same power (see Fig. 3b and 3d). Conversely, if power is very low, then the study will often run to completion without detecting the effect and the expected sample size will be close to the maximum sample size, which is greater than the fixed sample size. While we do see efficiency gains over fixed-sample designs, we find that permutation alpha spending is not as efficient as parametric alpha spending (see Fig. 3b and 2d); this is to be expected, as parametric methods are generally more efficient than non-parametric methods when their assumptions are met. For the statistical tests used in these simulations (*t*-max and cluster-based permutation tests), the assumptions of parametric alpha spending are *not* met, but we show the theoretical curve as a lower bound on the expected sample sizes we could in-principle anticipate from our permutation approach. Interestingly, cluster-based permutation tests are closer to the lower bound than is the *t*-max procedure, so the efficiency gains one can expect from permutation alpha spending may depend on the statistical test used, in contrast to parametric alpha spending (where relative efficiency depends only on the design parameters, e.g. spending function and number of interim analyses). However, gains should always be seen when the probability of early rejection is high.

In these simulations, we used a Pocock spending function, resulting in a more generous interim significance threshold than, for instance, an O'Brien-Fleming spending function, which spends more of the false positive rate later on during data collection. Please note, however, that the choice of spending function does affect the inflation factor and expected sample size of sequential designs, so these simulation results would vary between different spending functions. O'Brien-Fleming, for instance, tends to result in lower expected sample sizes than does Pocock, as it saves Type I error for interim sample sizes with more per-analysis statistical power. (However, it features lower alpha thresholds early on, which require more permutations to compute low enough *p*-values to reject – see *4.4 Limitations of the Present Approach.*)

4. Discussion

4.1. Use cases of sequential testing for neuroimaging

We anticipate at least several main use cases for permutation alpha spending in human neuroscience, as mentioned in the introduction:

1. A researcher does not have any reasonable means of running an a priori power analysis. Instead, they select a conservative maximum sample size they would be willing to collect, but they use permutation alpha spending to perform interim analyses throughout data collection while controlling their false positive rate. If the effect size ends up being large enough to detect with a smaller sample size, they will likely be able to stop data collection early.
2. A researcher has no reasonable means of running an a priori power analysis. Instead, they select an initial maximum sample size, and they perform a *conditional power analysis* by bootstrap after the first interim analysis, which they use to adjust their maximum sample size midway through data collection so as to achieve some desired power. In this way, they get to conduct a power analysis on the basis of some internal pilot data, but if the pilot data alone provide enough evidence to reject the null hypothesis, they need not collect any additional data.
3. A researcher has conducted a power analysis for a fixed-sample design, but they found that they need to collect a very large sample size, so they wish take advantage of the efficiency gains afforded by a sequential design. To this end, they use an inflation factor to convert the fixed-sample design for which they've already done a sample size calculation into similarly-powered sequential design with several interim analyses. Now, if they find there is already enough evidence to reject the null hypothesis at an interim analysis, they may stop data collection early.

In Example 2, the researcher could conduct a power analysis very similarly to how we have conducted the power analyses above. However, a *conditional* power analysis differs somewhat from an a priori power analysis in that it estimates the probability of a given design rejecting the null hypothesis *given the data already collected* (Spiegelhalter et al., 1986). We have included an (experimental) module for estimating conditional power by Bayesian bootstrap (Rubin, 1981) in our Python package, though only power analyses for one-sample (or paired-sample) tests are implemented at time of writing.

Also note that, if one adjusts the maximum sample size midway through, the alpha spending function must be adjusted accordingly to reflect the new design or the false positive rate will not be controlled (Lakens et al., 2021). One would similarly need to change the alpha spending function if, for example, they end up collecting fewer observations than their originally intended maximum sample size due to practical constraints. We provide an example of how to adjust one's alpha spending function mid-experiment on our package's GitHub and Zenodo repositories.

Even the researcher in the second example above may wish to run a conditional power analysis during their study if they estimated their sample size based on a previously published study, since that previous study may have overestimated its effect size as a result of low statistical power, publication bias, or selective reporting (e.g. reporting the effect sizes within significant clusters in a fMRI study) (Poldrack et al., 2017). The ability to adjust one's design on the basis of a conditional power analysis affords more options for navigating biases in the published literature while designing their own, well-powered study.

4.2. Power analyses, effect sizes, and sample size justification

While sequential designs offer an alternative to a priori sample size planning, it is often still desirable to perform some kind of power analysis to justify the prespecified maximum sample size (though it is not the only way to justify a sample size, as discussed at the end of this section). If it is possible to run an a priori power analysis for a fixed-sample design, then one can simply use an inflation factor to obtain the maximum sample size for their planned sequential design. If not, one can use a bootstrap conditional power calculation at an interim analysis and adjust the maximum sample size accordingly, or they can justify the maximum sample size based on a resource constraint (e.g. study budget, time constraint) or an implicit minimum effect size of interest.

As we have mentioned throughout this report, there is often not an obvious way to conduct a power analysis for the non-parametric tests we have discussed. Since the test statistics used (e.g. cluster mass) generally do not correspond to any standardized effect size, and the way the *p*-value is computed does not rely on distributional assumptions, usually the only way to estimate statistical power is by simulation – either by resampling from existing data or by using synthetic data.

In the resampling case, one conducts a power analysis based on the "effect size" implicit in some already collected data. Why implicit? Traditional power analyses are based on effect sizes (often, but not necessarily on a standardized scale) which importantly are not dependent on the size of the sample from which it was computed (Fritz et al., 2012; Kühberger et al., 2014). For permutation tests, there may not be a way of quantifying an effect size for the test statistic that is not dependent on the sample size; even if there is, defining an effect size of theoretical interest is usually challenging. For example, the cluster-mass statistic in a cluster-based permutation test (or the network-based statistic) cannot be called a standardized effect size, since the extent of clusters depends both on the sample size and on a number of computational parameters (Meyer et al., 2021; Sassenhagen and Draschkow, 2019). In the literature, some researchers have used bootstrap resampling from previously collected data to estimate the power of proposed sample-sizes (Ruzzoli et al., 2019), much in the same way we used bootstrap resampling to estimate power in our simulation here (see 2.3.2. Power/efficiency simulations). This approach implicitly incorporates effect size information from the original dataset, but it does so in a way that accounts for the dependency between the test statistic (e.g. cluster statistic) and the sample size by running a full permutation test on each bootstrap resample (at each sample size for which one wishes to estimate power).

However, any approach that involves estimating power based on the effect size measured in a previous study should take into consideration the intrinsic variability of effect size measurements. For instance, even effect size estimates do not become stable (i.e. likely to stay within a reasonable range of the estimate if more samples are added) until sample sizes are already quite large – larger, indeed, than would normally be required to reject the null hypothesis (Schönbrodt and Perugini, 2013). As such, it is usually inadvisable to conduct a conditional power analysis early on in data collection; it is better to wait until the interim sample size is somewhat substantial. In this setting, sequential analyses that utilize *conditional power analyses* (as in the third example in Section 4.1) can be particularly useful. One can use, for instance, the first 50 samples to estimate the conditional power of rejecting the null hypothesis by the time the maximum sample size is collected, given the data which has already been collected, by bootstrap resampling. However, if enough evidence has already been accrued to reject the null hypothe-

sis at this interim analysis, no further data collection needs to be done. This approach (while computationally burdensome) allows sample size calculations from "pilot" data that is internal to the study sample itself, rather than requiring a separate (costly) pilot sample. Bootstrap conditional power analyses (using the procedure described in 2.3.2) are implemented in the "power" module of the niseq package for all one-sample and paired-sample tests in the package, and we hope to add support for independent sample and correlational designs too.

Another approach is to use synthetic data, such as a statistical map generated from a meta-analysis (Yarkoni et al., 2011), a predictive model (Dockès et al., 2020), or a formal theoretical/computational model, and compute a power analysis by simulation (adding noise to the predicted effect map on each simulation).

Or alternatively, if the goal of the study is the fit a predictive model that predicts a single behavioral outcome from whole-brain patterns (Kragel et al., 2018), rather than running a separate univariate test on every voxel in the brain, then a better approach may be to specify a *minimum effect size of interest*, which is not dependent on a noisy estimate from previous data but on some metric of (e.g. clinical) importance.

And lastly, it is also acceptable to justify one's maximum sample size in a sequential design based on a resource constraint rather than on a power analysis (Lakens, 2022a). For example, "We used a sequential design with a maximum sample size of 80, after which the cost of collecting subjects would outstrip the budget allocated for this research project." Alternatively, one might consider a specification of a maximum sample size to implicitly reveal the minimum effect size of interest – as in, an effect is not large enough to be of interest unless it is large enough to be detected with fewer than n subjects. (For instance, if more than n subjects are required to detect an effect, then the researcher would be unlikely to pursue that effect in future work, choosing to focus on something more cost-efficient to study.) In any event, it is not always strictly necessary to perform a power analyses for sample size justification depending on the goals of the study. (If the goal is to defend a null result, for instance, then it may be necessary, but perhaps not for an exploratory study.) In any event, researchers should transparently report how they determined their maximum sample size, even if that sample size is arbitrary.

4.3. Stopping for futility

In the simulations we feature in this article, we assume that data collection continues until the specified maximum sample size unless the null hypothesis is rejected at an interim analysis. Another option, however, is to run a conditional power analysis after the first (or any/each) interim analysis to estimate the probability of rejecting the null hypothesis if the design is run to completion. If this power falls below some (predetermined) threshold, then one might choose to stop the study for futility, rather than waste resources by continuing (Lan and Trost, 1997). In this case, one might achieve even greater sample size savings than we have described above.

If conditional power is estimated either from an effect size computed from the data that has already been collected or by resampling from that data (as implemented in niseq's "power" module), then the resulting power estimate is interpreted as the probability of detecting *any effect at all* if the current trend in the data continues. This approach is relatively common in the clinical trial literature (Lachin, 2005; Snapinn et al., 2006), and we think it is reasonable for exploratory neuroimaging studies. However, if a researcher aims to detect a *specific* effect – for instance, performance of a whole-brain predictive model that is high enough to be clinically useful – then it would make more sense to compute conditional power for that smallest effect size of interest. Similarly, if one has a specific hypothesis related to some region-of-interest, it may not make sense to run a power analysis as if one cared equally about the whole brain.

However, it is important to note that including a stopping rule for futility in one's design may affect the statistical power of the design. See work by Lakens and colleagues for a somewhat more thorough discussion (Albers and Lakens, 2018; Lakens, 2014; Lakens et al., 2021; Lakens and Evers, 2014). Moreover, the effective false positive rate will end up below the specified significance level, since the study is terminated before all Type I error is "spent" (Lachin, 2005). Whether one can "reclaim" the lost Type I error is an active area of research in the application of parametric alpha spending (Snapinn et al., 2006), but the extent to which solutions used in the parametric setting extend to the permutation case requires more research. In principle, it is possible to estimate these effects by simulation, but this would require substantial computation (since power analyses would be nested within a larger power analysis, and the above simulations already took a great deal of compute time). We hope to develop this area more thoroughly in the future, as we believe the principled use of futility stopping rules in sequential designs stands to greatly reduce the cost of neuroimaging research.

4.4. Reporting the results of a sequential analysis

There are not yet uniformly agreed-upon standards for reporting the results of sequential analyses. However, sequential analyses should report, minimally, the spending function used, the time of all interim analyses performed, the adjusted significance thresholds and value of the alpha spending function at each analysis, and the p-values observed at each analysis. In the case of t-max and other max-type procedure, as well as with threshold-free cluster enhancement, in which each voxel is assigned a p-value, we suggest reporting the smallest p-value obtained at each interim analysis and reporting full results for the time at which data collection was stopped. See the parametric alpha spending tutorial by Lakens and colleagues for more discussion (Lakens et al., 2021).

Note that it has been suggested that sequential analyses should be pre-registered for the sake of transparency (Lakens, 2014), which may be helpful in providing evidence that one did not change their design parameters in the analysis stage. Since the number and timing of interim analyses does not need to be determined a priori for the alpha spending procedure to control the false positive rate, then it is sufficient to specify a statistical test, an alpha spending function, a maximum sample size, and whether a stopping rule for futility will be used. However, it is also helpful to specify tentative interim analysis times, even though these may be altered without issue later on.

4.5. Multiple testing in fixed-sample and sequential designs

As illustrated in our simulations, permutation alpha spending can control the familywise false positive rate across an arbitrary number of statistical tests conducted in parallel when combined with any existing permutation-based multiple comparisons correction. Here, we have highlighted interoperability with t-max and cluster-based corrections, which are paradigmatic of the two primary approaches to multiple comparisons correction in neuroimaging: voxel-level and cluster-level. In both cases, permutation alpha spending works by finding a rejection threshold at each interim analysis that controls the familywise error rate (across all sequential analyses) instead of the error rate of any one test (voxel).

In voxel-level inference, control of the familywise error rate is achieved by adjusting the all voxels' rejection threshold such that the probability of falsely rejecting the null hypothesis at *any* voxel is contained below the target error rate. A simple approach to compute the adjusted threshold is the well-known Bonferroni correction, in which the required adjustment is purely a function of the number of tests performed, as the correction assumes v independent comparisons (Miller, 2012). In the event that the v statistical tests are correlated (as in the case almost all neuroimaging settings), the actual "effective" number of comparisons can be thought of as less than v; consequently, neuroimaging researchers have long favored corrections that account for autocorrelation across space, time, and/or frequency. While there are (still

popular) parametric approaches to this problem (Friston et al., 1994), which make some assumptions about the autocorrelation structure of the data, it has been known for decades that permutation-based corrections have the potential to outperform parametric ones by implicitly estimating the dependency between tests from the data (Holmes et al., 1996). The popular *t*-max correction, for instance, is a permutation test that uses the *maximum t*-statistic observed across *all* voxels in the volume of interest on each permutation to construct the permutation null distribution (obtained by randomly shuffling the observations); individual voxel *t*-statistics are then compared to the maximum *t* null distribution rather than the null distribution for that voxel, and "corrected" *p*-values are assigned as the percentage rank of a voxel's observed test statistic in that null distribution (Nichols and Holmes, 2002). This ensures that the probability of *any* voxel's corrected *p*-value falling below the specified significance level at the desired familywise error rate. In the sequential case, the joint null distribution of the maximum *t* statistic across all interim analyses is used to find adjusted significance levels as described in Section 2.2; thus, each *p*-value is compared to the alpha-spending-adjusted threshold, which additionally accounts for the correlation between tests statistics across sequential analyses.

In a typical cluster-level approach, a clustering threshold is determined a priori (and somewhat arbitrarily), often set to the value at which an individual voxel would be labelled "significant" without a correction for multiple comparisons. Then, within each cluster of above-threshold contiguous voxels, the voxel-level test statistics are aggregated (usually summed, or in the case of the network-based statistic, which is essentially a cluster-based test, the number of included edges are counted) to produce a cluster mass statistic. Instead of comparing each voxel to a null distribution, a *p*-value is assigned to each as the percentile rank of the observed cluster mass statistic compared to a null distribution computed across many random shuffles of the data (Maris and Oostenveld, 2007). By abstracting inference to the cluster level in this way, one tends to gain statistical power by reducing what would be many voxel-level tests to, effectively, a single test, one loses the ability to make precise inferences about the voxel-level locus of the effect (Sassenhagen and Draschkow, 2019). Again, the only modification in the sequential case is that each clusters' *p*-value is compared to the adjusted significance threshold computed using the alpha spending approach instead of the nominal significance level.

In this way, sequential alpha spending can be combined with any permutation-based approach to controlling the familywise error rate that yields "corrected" *p*-values computed as a percentage rank in a permutation null distribution. The resulting rejection thresholds control the familywise error rate across all interim analyses (i.e. the probability of getting any false positives at all throughout the whole procedure). See Section 4.5, however, for limitations of using the sequential approach in conjunction with multiple comparisons correction.

4.6. Limitations of the present approach

The present approach has a number of limitations that are worth discussing.

If one stops data collection for a neuroimaging study as soon as one accrues enough evidence to reject the null hypothesis on the basis of a cluster (e.g. in a cluster-level test) or a voxel (e.g. in a *t*-max procedure), one might miss other, more weakly activated clusters or voxels elsewhere in the data. In principle, this is also a limitation of fixed-sample designs with low statistical power, so a well-powered sequential design, we think, is still usually preferable to arbitrary or heuristic sample size determination. Moreover, unlike a fixed-sample design in which the sample size is (supposed to be) set in stone at the time of analysis, a sequential design allows one to keep collecting data even after the null hypothesis is rejected. In this way, if the researcher sees a cluster that was "trending toward significance" at the time they rejected the null based on a stronger cluster, they can continue to collect data until they reach their prespecified maximum sample size before concluding that they failed to reject the null. This possibility is illustrated in the fMRI example below.

Further, when a sequential design stops at an interim analysis, there is a risk that the test statistic crossed the rejection threshold because, due to random variation, the effect size was overestimated at that interim sample size. This does not inflate the false positive rate, as this risk is accounted for in the null distribution, but may result in biased effect size estimates for sequential designs. This may, however, be less of a problem for permutation alpha spending than for parametric alpha spending, because the exact numerical value of the test statistic used in the permutation test (e.g. cluster mass) is usually not directly of interest in neuroimaging studies. Moreover, this bias is washed out in meta-analyses, since effect sizes measured in studies that are terminated early are balanced out by those that ran to completion (Schönbrodt et al., 2017). This fact underscores the importance of sharing unthresholded statistical maps from neuroimaging studies, which can be hosted on platforms such as NeuroVault to facilitate future metanalyses (Gorgolewski et al., 2016).

Both of the above issues can also be optionally alleviated, if a researcher wishes, by continuing to collect data past the interim analysis at which the null hypothesis is rejected. There is nothing stopping data collection after the null hypothesis has already been rejected in the interest of yielding better estimates of the effect of interest – or, better yet, estimating that effect in an independent sample.

Finally, with any permutation-based method, one requires a sufficiently large sample size for valid inference. For instance, the number of possible permutations for a one-sample permutation test with $n = 5$ observations is $2^5 = 32$, which is far too small; the lowest p-value one could possibly compute with that few permutations is $1/32 = 0.031$. Thus, an interim analysis at $n = 5$ has no chance of rejecting the null hypothesis if the adjusted significance threshold for that analysis is lower than 0.031 and still a substantially reduced chance otherwise. Researchers should ensure that the number of possible permutations at their smallest interim sample size is sufficient to reject the null hypothesis at the adjusted significance level determined by their alpha spending function.

4.7. The Bayesian alternative

Frequentist approaches multiple testing and to sequential hypothesis testing, as featured here, are often compared to approaches that use Bayes factors. It is often claimed that Bayesian statistical approaches suffer from neither a multiple comparisons problem nor a problem with optional stopping. Since our approach applies (frequentist) sequential tests in a multiple testing context, it is worth addressing both these claims separately, as the error rate properties of Bayesian approaches are more nuanced than some proponents in applied fields argue.

Discussions of the multiple comparisons problem can be obfuscated by differences in the frequentist and Bayesian philosophies. Bayesian statistics does not aim to make binary decisions; a Bayesian never "rejects" a hypothesis on the basis of new data, but merely adjusts the posterior likelihood assigned to each possible hypothesis. In that trivial sense, the idealized Bayesian never suffers from inflated false positive rates; one cannot make an error when one is not making any hard decisions. However, one can still consider the frequentist (error rate) properties of a Bayesian estimator, and indeed such evaluations are considered useful performance metrics by practicing Bayesian statisticians (Gelman and Carlin, 2014).

In this setting, the answer to whether Bayesian approaches suffer from a multiple comparisons problem amounts to "it depends." For instance, the authors of one paper frequently cited to support to claim that there is no Bayesian multiple comparisons problem (Gelman et al., 2012), also make clear that the desirable error rate properties of Bayesian approaches only apply in the context of hierarchical models that account for the interdependency between the measured variables (e.g. voxels) explicitly, rather than to approaches that would treat each variable as a separate test (Gelman and Tuerlinckx, 2000). In the context

of neuroimaging, this amounts to a spatially autoregressive model that biases the estimates at neighboring voxels toward each other, thus accounting for multiple comparisons by shrinking the effect size estimates – in contrast to a frequentist approach which accounts for multiple comparisons by increasing the width of the confidence interval (i.e. adjusting the *p*-value) (Lindquist et al., 2009). In other words, replacing the mass-univariate *t*-test with a mass-univariate Bayes factor analysis would not solve the multiple comparisons problem in neuroimaging, but using an autoregressive model with regularizing priors might (Harrison and Green, 2010). Since Bayes factors do not lend themselves to an obvious correction for multiple comparisons (outside of a frequentist framework), they tend to be inapplicable to the mass-univariate voxel-level inference employed in many neuroimaging studies. Moreover, there is not, to our knowledge, a Bayes factor approach to cluster-level inference that does not depend on strong parametric assumptions such as those entailed by random field theory (Friston et al., 2002).

Similarly, it is often claimed that Bayes factors allow for optional stopping. It is true, in many cases, that the interpretation of a Bayes factor as a measure of relative evidence for competing hypotheses is not affected by the stopping rule used during sequential data collection (Rouder, 2014); but also see (de Heide and Grünwald, 2021). However, Bayes factors do not provide error rate guarantees, and inflated false positive rates (in the frequentist sense) can occur when using a Bayes factor threshold as a stopping rule for sequential hypothesis testing (Schönbrodt et al., 2017); thus, the stopping rule (threshold) needs to be calibrated by simulation to ensure reasonable error rates (Schönbrodt and Wagenmakers, 2018). Thus, Bayes factors have the advantage that they are interpretable measures of evidence, unlike *p*-values (Lakens, 2022b), but they do not provide the strong error rate guarantees as do frequentist approaches.

Interestingly, the fact that permutation alpha spending is valid with arbitrary test statistics open the door to combining frequentist and Bayesian sequential test procedures. If a Bayes factor is employed as a test statistic in a frequentist sequential test, then strong error rate guarantees will still hold, but the final observed Bayes factor will serve as an interpretable measure of evidence, unbiased by the sequential stopping rule (Rouder, 2014), unlike a typical effect size measure at the end of a sequential procedure (see Section 4.7). In a neuroimaging context, this would entail replacing the voxel-level *t*-test with a Bayes factor analysis, and using either a sequential cluster test or sequential max-type correction as described in this report.

4.8. The niseq package

We introduce a Python package, called "niseq," geared toward applying the permutation alpha spending approach to M/EEG, fMRI, and connectivity data (although the sequential test implementations could be used on other datatypes as well). At time of writing, the package includes sequential implementations of *t*-max (as well as other max-type tests, e.g. *F*-max and *r*-max) (Nichols and Holmes, 2002), cluster-based permutation tests (Maris and Oostenveld, 2007), and threshold-free cluster enhancement (Smith and Nichols, 2009). Since the network-based statistic (NBS) (Zalesky et al., 2010) is simply a cluster-based permutation test applied to subnetworks (clusters) of connected graph edges, a sequential version of this procedure can also be easily applied with niseq. In the package repository (see Data and Code Availability), numerous tutorial examples are provided for all of these tests using M/EEG, fMRI, and connectivity data. (Niseq also contains an in-development "niseq.power" module for running prospective and conditional power analyses by bootstrap, but it is currently only compatible with one-tailed tests available in the package; at time of writing, it should be considered experimental. The focus of the package, at the moment, is on permutation alpha spending itself.)

To minimize friction for users of existing packages in Python's human neuroscience ecosystem, we attempted to mirror the MNE-Python API as much as possible (Gramfort et al., 2014). We chose to mirror the MNE API, rather than another package in the neuroimaging ecosystem, since MNE's statistics functions act directly on numerical arrays, rather than on M/EEG- or MRI-specific data structures (which usually contain such arrays internally). Thus, the same API can be used for many datatypes with minimal hassle. For users familiar with MNE, this means that analysis code will only have to be minimally edited (see Fig. 4). For fMRI-oriented users coming from packages in the nipy ecosystem (Brett et al., 2009), data arrays will have to be pulled out of MRI specific objects, as in the usage example below.

Imagine a researcher is running an fMRI experiment with the aim of localizing brain activity associated with the calculation task described by Pinel and colleagues (Pinel et al., 2007). She has received an institutional grant for this study, which can pay for her to collect data from up to 80 subjects; however, if possible, she would like to collect fewer observations to conserve funds. Consequently, she decides to use a sequential design with a maximum sample size of 80, and she plans to run an interim analysis every 10 subjects.

She is used to analyzing her fMRI data using nilearn (nilearn.github.io). Fortunately, she can conduct her first-level analysis in nilearn as usual to extract statistical (Z-) maps for each subject, using the GLM contrast described by Pinel and colleagues. Her workflow would only change at the second-level (i.e. group) analysis stage, in which she would pull out the data arrays from the nibabel and subject them to a permutation test in the niseq package. She uses a one-sample cluster-based permutation test with a clustering threshold of $t = 3$, and a significance level of $\alpha = 0.01$ to compare her statistical maps to a null hypothesis in which there is no task-related activation above baseline. For alpha spending, she uses nilearn's default spending function, which is linear. She uses 5000 permutations, so the lowest *p*-value she can find at any interim analysis is $1/5000 = 0.0002$, ensuring it is possible to reject the null at the adjusted significance threshold at the first interim analysis ($\alpha = 0.00125$). This sequential test procedure can be executed using the "sequential_cluster_test_1samp" function in niseq.

On her first interim analysis at $n = 10$, the adjusted significance threshold is just the value of the spending function ($\alpha = 0.00125$). Though the smallest cluster *p*-value she computes is below the nominal significance level of 0.01 at $p = 0.008$, it is still higher than the adjusted significance level for the interim analysis and she cannot reject the null hypothesis. Consequently, she collects another 10 subjects, and she runs another interim analysis. The value of the spending function at $n = 20$ is 0.0025, so the permutation procedure described in this paper is used to find the adjusted significance threshold which contains the cumulative false positive rate to 0.0025 ($\alpha = 0.0024$). She finds that her largest cluster now has a *p*-value below the adjusted threshold for this interim analysis (see Fig. 5). However, she sees a cluster in parietal cortex that was nominally significant at $\alpha = 0.01$, but still not significant given the alpha spending correction. She suspects that this might represent true brain activity (though she currently does not have enough evidence to reject the null hypothesis of no brain activity), so she collects another 10 subjects. At $n = 30$, the permutation procedure finds that an adjusted significance threshold of $\alpha = 0.002$ is needed to contain the cumulative false positive rate to the spending function's value of 0.00375. Running another interim analysis, she finds she now has enough evidence to reject the null hypothesis on the basis of a cluster containing the aforementioned parietal location as well. Results of this analysis procedure applied to the calculation localizer data in the Brainomics dataset (Papadopoulos Orfanos et al., 2017) can be seen in Fig. 5, and code to reproduce the sequential analysis workflow described in this paragraph can be found at https://doi.org/10.5281/zenodo.7926956.

4.9. Outlook

We believe that sequential designs can be a valuable tool as cognitive neuroscientists continue their efforts to improve the statistical power of neuroimaging studies, while balancing costs. Sequential designs provide multiple alternative paths to principled sample size determination in the

```python
from mne.stats import permutation_cluster_1samp_test
import mne

# load MNE evoked response objects from file
evokeds = mne.read_evokeds('your-data.fif')
# pull out data arrays from MNE objects
X = np.stack([evo.get_data() for evo in evokeds], axis = 0)
X = np.transpose(X, (0, 2, 1)) # observations x time x chans
# get channel adjacency matrix
adj, _ = mne.channels.find_ch_adjacency(evokeds[0].info, "eeg")

# and finally, run cluster test
t_map, clusts, clust_ps, H0 = permutation_cluster_1samp_test(
    X,
    threshold = 3, # t-val threshold for clustering
    adjacency = adj
)
```

```python
from niseq import sequential_cluster_test_1samp
import mne

# load MNE evoked response objects from file
evokeds = mne.read_evokeds('your-data.fif')
# pull out data arrays from MNE objects
X = np.stack([evo.get_data() for evo in evokeds], axis = 0)
X = np.transpose(X, (0, 2, 1)) # observations x time x chans
# get channel adjacency matrix
adj, _ = mne.channels.find_ch_adjacency(evokeds[0].info, "eeg")

# and finally, run cluster test
n = X.shape[0] # sample size
looks, ps, adj_alphas, spending = sequential_cluster_test_1samp(
    X,
    look_times = [n], # One analysis
    n_max = n, # at the max sample size...
    threshold = 3, # ...is equivalent to a fixed-sample test.
    adjacency = adj
)
t_map, clusts, clust_ps, H0 = looks[n]
```

Fig. 4. Comparison of MNE-Python and niseq APIs. These code samples run identical cluster-based permutation tests on EEG data using MNE-Python (left) and using niseq (right), demonstrating the equivalence between the two APIs. However, the niseq version can be modified to run a sequential test by changing the "look_times" and "n_max" arguments. Niseq's statistics functions will return a dictionary containing results (in the same format as MNE) for each interim analysis, the lowest cluster/voxel *p*-value at each analysis, as well as the adjusted alphas and value of the alpha spending function at each interim analysis.

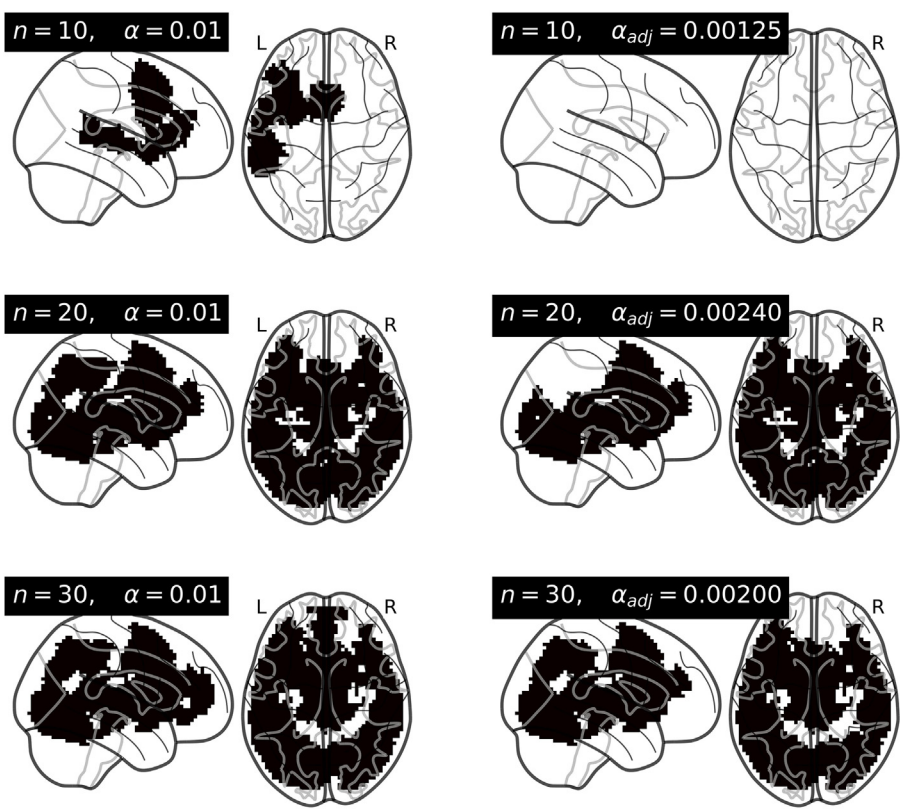

Fig. 5. Example results of a sequential cluster-based permutation test with three interim analyses. Clusters for which the *p*-value is smaller than a given significance level are shaded in black. For each interim analysis (row), uncorrected significance thresholds are used on the left and alpha spending corrected thresholds are used on the right. Tutorial code for reproducing this figure (and walking through a sequential analysis workflow for fMRI) using data from the Brainomics dataset can be found at https://doi.org/10.5281/zenodo.7926956.

event that conventional, a priori power analyses are difficult to perform. Moreover, even in the event that one can easily perform a power analysis for a well-specified effect of interest, highly-powered studies (e.g. confirmatory trials) stand to benefit greatly from the efficiency advantages of sequential designs. Indeed, when each subject costs hundreds or even thousands of dollars to run, as in an fMRI study, a greater than 30% reduction in expected sample size without sacrificing statistical power (see Fig. 3) can free up valuable resources for cognitive neuroimaging labs and their funding agencies. We hope that our permutation-based approach to sequential analysis proposed in this article, and our accompanying Python package, empower cognitive neuroscience researchers to conduct more efficient studies.

Declaration of Competing Interest

The authors declare no conflict of interest.

Credit authorship contribution statement

John P. Veillette: Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Project administration, Software, Validation, Visualization, Writing – original draft, Writing – review & editing. **Letitia Ho:** Data curation, Formal analysis, Investigation, Validation, Writing – review & editing. **Howard C. Nusbaum:** Funding acquisition, Resources, Supervision, Writing – review & editing.

Data availability

Our implementations of permutation alpha spending for cluster-based permutation tests, threshold-free cluster enhancement, t-max, F-max, r-max, and the network-based statistic are contained in our user-friendly Python package niseq, which can be installed from the Python package index (PyPI). Documentation is hosted on Read the Docs (http://niseq.readthedocs.io/). Source code, as well as worked examples using the package on EEG and fMRI data in conjunction with the MNE-Python and nilearn packages, are available on GitHub (https://github.com/john-veillette/niseq) and permanently archived on Zenodo. The most recent release as of writing (v0.0.2) is available at https://doi.org/10.5281/zenodo.7526535 and the current release is always archived at https://doi.org/10.5281/zenodo.7517285.

The code used for the simulations featured in this article, as well as the results of those simulations and a record of the simulation parameters, are available separately on GitHub (https://github.com/john-veillette/niseq-simulations) and are permanently archived on Zenodo (https://doi.org/10.5281/zenodo.7666443).

The data used for simulations was originally taken from the ERP CORE repository on the Open Science Framework (https://doi.org/10.18115/D5JW4R). However, we exported the preprocessed difference waves into a file format that could be easily loaded with MNE-Python, which we provide for convenience in the same Zenodo archive as our simulation code.

Funding

This work was supported by the National Science Foundation [grant numbers GRFP DGE 1746045, NCS 2024923, and NCS 1835181].

Acknowledgements

This work was completed in part with resources provided by the University of Chicago's Research Computing Center.

References

Albers, C., Lakens, D., 2018. When power analyses based on pilot data are biased: inaccurate effect size estimators and follow-up bias. J. Exp. Soc. Psychol. 74, 187–195. doi:10.1016/j.jesp.2017.09.004.

Armitage, P., McPherson, C.K., Rowe, B.C., 1969. Repeated significance tests on accumulating data. J. R. Stat. Soc. Ser. Gen. 132, 235–244. doi:10.2307/2343787.

Brett, M., Taylor, J., Burns, C., Millman, K., Perez, F., Roche, A., Thirion, B., D'Esposito, M., 2009. NIPY: an open library and development framework for FMRI data analysis. In: NeuroImage, Organization for Human Brain Mapping 2009 Annual Meeting, 47, p. S196. doi:10.1016/S1053-8119(09)72223-2.

Button, K.S., Ioannidis, J.P.A., Mokrysz, C., Nosek, B.A., Flint, J., Robinson, E.S.J., Munafò, M.R., 2013. Power failure: why small sample size undermines the reliability of neuroscience. Nat. Rev. Neurosci. 14, 365–376. doi:10.1038/nrn3475.

de Heide, R., Grünwald, P.D., 2021. Why optional stopping can be a problem for Bayesians. Psychon. Bull. Rev. 28, 795–812. doi:10.3758/s13423-020-01803-x.

de Vrieze, J., 2021. Large survey finds questionable research practices are common. Science 373, 265. doi:10.1126/science.373.6552.265, 265.

Dockès, J., Poldrack, R.A., Primet, R., Gözükan, H., Yarkoni, T., Suchanek, F., Thirion, B., Varoquaux, G., 2020. NeuroQuery, comprehensive meta-analysis of human brain mapping. Elife 9, e53385. https://doi.org/10.7554/eLife.53385.

Dodge, H.F., Romig, H.G., 1929. A method of sampling inspection. Bell Syst. Tech. J. 8, 613–631. doi:10.1002/j.1538-7305.1929.tb01240.x.

Durnez, J., Degryse, J., Moerkerke, B., Seurinck, R., Sochat, V., Poldrack, R., Nichols, T., 2016. Power and sample size calculations for fMRI studies based on the prevalence of active peak. https://doi.org/10.1101/049429.

Friston, K.J., Holmes, A.P., Worsley, K.J., Poline, J.P., Frith, C.D., Frackowiak, R.S.J., 1994. Statistical parametric maps in functional imaging: a general linear approach. Hum. Brain Mapp 2, 189–210. doi:10.1002/hbm.460020402.

Friston, K.J., Penny, W., Phillips, C., Kiebel, S., Hinton, G., Ashburner, J., 2002. Classical and bayesian inference in neuroimaging: theory. Neuroimage 16, 465–483. doi:10.1006/nimg.2002.1090.

Fritz, C.O., Morris, P.E., Richler, J.J., 2012. Effect size estimates: current use, calculations, and interpretation. J. Exp. Psychol. Gen. 141, 2–18. doi:10.1037/a0024338.

Gelman, A., Carlin, J., 2014. Beyond power calculations: assessing type S (Sign) and type M (Magnitude) errors. Perspect. Psychol. Sci. 9, 641–651. doi:10.1177/1745691614551642.

Gelman, A., Hill, J., Yajima, M., 2012. Why we (Usually) don't have to worry about multiple comparisons. J. Res. Educ. Eff. 5, 189–211. doi:10.1080/19345747.2011.618213.

Gelman, A., Tuerlinckx, F., 2000. Type S error rates for classical and Bayesian single and multiple comparison procedures. Comput. Stat. 15, 373–390. doi:10.1007/s001800000040.

Glimm, E., Maurer, W., Bretz, F., 2010. Hierarchical testing of multiple endpoints in group-sequential trials. Stat. Med. 29, 219–228. doi:10.1002/sim.3748.

Gorgolewski, K.J., Varoquaux, G., Rivera, G., Schwartz, Y., Sochat, V.V., Ghosh, S.S., Maumet, C., Nichols, T.E., Poline, J.B., Yarkoni, T., Margulies, D.S., Poldrack, R.A., 2016. NeuroVault.org: a repository for sharing unthresholded statistical maps, parcellations, and atlases of the human brain. NeuroImage, Sharing the wealth: brain Imaging Repositories in 2015 124, 1242–1244. doi:10.1016/j.neuroimage.2015.04.016.

Gramfort, A., Luessi, M., Larson, E., Engemann, D.A., Strohmeier, D., Brodbeck, C., Parkkonen, L., Hämäläinen, M.S., 2014. MNE software for processing MEG and EEG data. Neuroimage 86, 446–460. doi:10.1016/j.neuroimage.2013.10.027.

Guo, Q., Thabane, L., Hall, G., McKinnon, M., Goeree, R., Pullenayegum, E., 2014. A systematic review of the reporting of sample size calculations and corresponding data components in observational functional magnetic resonance imaging studies. Neuroimage 86, 172–181. doi:10.1016/j.neuroimage.2013.08.012.

Harrison, L.M., Green, G.G.R., 2010. A Bayesian spatiotemporal model for very large data sets. Neuroimage 50, 1126–1141. doi:10.1016/j.neuroimage.2009.12.042.

Holmes, A.P., Blair, R.C., Watson, J.D., Ford, I., 1996. Nonparametric analysis of statistic images from functional mapping experiments. J. Cereb. Blood Flow Metab. Off. J. Int. Soc. Cereb. Blood Flow Metab. 16, 7–22. doi:10.1097/00004647-199601000-00002.

Ioannidis, J.P.A., 2005. Why most published research findings are false. PLOS Med 2, e124. doi:10.1371/journal.pmed.0020124.

John, L.K., Loewenstein, G., Prelec, D., 2012. Measuring the prevalence of questionable research practices with incentives for truth telling. Psychol. Sci. 23, 524–532. doi:10.1177/0956797611430953.

Joyce, K.E., Hayasaka, S., 2012. Development of PowerMap: a software package for statistical power calculation in neuroimaging studies. Neuroinformatics 10, 351–365. doi:10.1007/s12021-012-9152-3.

Kappenman, E.S., Farrens, J.L., Zhang, W., Stewart, A.X., Luck, S.J., 2021. ERP CORE: an open resource for human event-related potential research. Neuroimage 225, 117465. doi:10.1016/j.neuroimage.2020.117465.

Kosorok, M.R., Yuanjun, S., DeMets, D.L., 2004. Design and analysis of group sequential clinical trials with multiple primary endpoints. Biometrics 60, 134–145. doi:10.1111/j.0006-341X.2004.00146.x.

Kragel, P.A., Koban, L., Barrett, L.F., Wager, T.D., 2018. Representation, pattern information, and brain signatures: from neurons to neuroimaging. Neuron 99, 257–273. doi:10.1016/j.neuron.2018.06.009.

Kühberger, A., Fritz, A., Scherndl, T., 2014. Publication bias in psychology: a diagnosis based on the correlation between effect size and sample size. PLoS ONE 9, e105825. doi:10.1371/journal.pone.0105825.

Lachin, J.M., 2005. A review of methods for futility stopping based on conditional power. Stat. Med. 24, 2747–2764. doi:10.1002/sim.2151.

Lakens, D., 2022a. Sample size justification. Collabra Psychol 8, 33267. doi:10.1525/collabra.33267.

Lakens, D., 2022b. Why P values are not measures of evidence. Trends Ecol. Evol. 37, 289–290. doi:10.1016/j.tree.2021.12.006.

Lakens, D., 2014. Performing high-powered studies efficiently with sequential analyses. Eur. J. Soc. Psychol. 44, 701–710. doi:10.1002/ejsp.2023.

Lakens, D., Evers, E.R.K., 2014. Sailing from the seas of chaos into the corridor of stability: practical recommendations to increase the informational value of studies. Perspect. Psychol. Sci. 9, 278–292. doi:10.1177/1745691614528520.

Lakens, D., Pahlke, F., Wassmer, G., 2021. Group sequential designs: a tutorial. https://doi.org/10.31234/osf.io/x4azm.

Lan, G., DeMets, D., 1983. Discrete sequential boundaries for clinical trials. Biometrika 70, 659–663. doi:10.1093/biomet/70.3.659.

Lan, K.K.G., Trost, D.C., 1997. Estimation of Parameters and Sample Size Re-Estimation. Biopharmaceutical Section American Statistical Association, pp. 48–51.

Lancaster, T., 2003. A note on bootstraps and robustness. Brown Univ. Dep. Econ. Work. Pap. doi:10.2139/ssrn.896764.

Larson, M.J., Carbine, K.A., 2017. Sample size calculations in human electrophysiology (EEG and ERP) studies: a systematic review and recommendations for increased rigor. Int. J. Psychophysiol., Rigor Replicat.: Towards Improved Best Pract. Psychophysiol. Res. 111, 33–41. doi:10.1016/j.ijpsycho.2016.06.015.

Lindquist, M.A., Gelman, A., 2009. Correlations and multiple comparisons in functional imaging: a statistical perspective (Commentary on Vul et al., 2009). Perspect. Psychol. Sci. 4, 310–313. doi:10.1111/j.1745-6924.2009.01130.x.

Marek, S., Tervo-Clemmens, B., Calabro, F.J., Montez, D.F., Kay, B.P., Hatoum, A.S., Donohue, M.R., Foran, W., Miller, R.L., Hendrickson, T.J., Malone, S.M., Kandala, S., Feczko, E., Miranda-Dominguez, O., Graham, A.M., Earl, E.A., Perrone, A.J., Cordova, M., Doyle, O., Moore, L.A., Conan, G.M., Uriarte, J., Snider, K., Lynch, B.J., Wilgenbusch, J.C., Pengo, T., Tam, A., Chen, J., Newbold, D.J., Zheng, A., Seider, N.A., Van, A.N., Metoki, A., Chauvin, R.J., Laumann, T.O., Greene, D.J., Petersen, S.E., Garavan, H., Thompson, W.K., Nichols, T.E., Yeo, B.T.T., Barch, D.M., Luna, B., Fair, D.A., Dosenbach, N.U.F., 2022. Reproducible brain-wide association studies require thousands of individuals. Nature 603, 654–660. doi:10.1038/s41586-022-04492-9.

Maris, E., Oostenveld, R., 2007. Nonparametric statistical testing of EEG- and MEG-data. J. Neurosci. Methods 164, 177–190. doi:10.1016/j.jneumeth.2007.03.024.

Meyer, M., Lamers, D., Kayhan, E., Hunnius, S., Oostenveld, R., 2021. Enhancing reproducibility in developmental EEG research: BIDS, cluster-based permutation tests, and effect sizes. Dev. Cogn. Neurosci. 52, 101036. doi:10.1016/j.dcn.2021.101036.

Miller, R.G.J., 2012. Simultaneous Statistical Inference. Springer Science & Business Media.

Mumford, J.A., Nichols, T.E., 2008. Power calculation for group fMRI studies accounting for arbitrary design and temporal autocorrelation. Neuroimage 39, 261–268. doi:10.1016/j.neuroimage.2007.07.061.

Nichols, T.E., Holmes, A.P., 2002. Nonparametric permutation tests for functional neuroimaging: a primer with examples. Hum. Brain Mapp 15, 1–25. doi:10.1002/hbm.1058.

Papadopoulos Orfanos, D., Michel, V., Schwartz, Y., Pinel, P., Moreno, A., Le Bihan, D., Frouin, V., 2017. The Brainomics/Localizer database. NeuroImage, Data Sharing Part II 144, 309–314. doi:10.1016/j.neuroimage.2015.09.052.

Pinel, P., Thirion, B., Meriaux, S., Jobert, A., Serres, J., Le Bihan, D., Poline, J.B., Dehaene, S., 2007. Fast reproducible identification and large-scale databasing of individual functional cognitive networks. BMC Neurosci. 8, 91. doi:10.1186/1471-2202-8-91.

Poldrack, R.A., Baker, C.I., Durnez, J., Gorgolewski, K.J., Matthews, P.M., Munafò, M.R., Nichols, T.E., Poline, J.B., Vul, E., Yarkoni, T., 2017. Scanning the horizon: towards transparent and reproducible neuroimaging research. Nat. Rev. Neurosci. 18, 115–126. doi:10.1038/nrn.2016.167.

Rosenberg, M.D., Finn, E.S., 2022. How to establish robust brain–behavior relationships without thousands of individuals. Nat. Neurosci. 25, 835–837. doi:10.1038/s41593-022-01110-9.

Rouder, J.N., 2014. Optional stopping: no problem for Bayesians. Psychon. Bull. Rev. 21, 301–308. doi:10.3758/s13423-014-0595-4.

Rubin, D.B., 1981. The Bayesian bootstrap. Ann. Stat. 9, 130–134. doi:10.1214/aos/1176345338.

Ruzzoli, M., Torralba, M., Morís Fernández, L., Soto-Faraco, S., 2019. The relevance of alpha phase in human perception. Cortex 120, 249–268. doi:10.1016/j.cortex.2019.05.012.

Sassenhagen, J., Draschkow, D., 2019. Cluster-based permutation tests of MEG/EEG data do not establish significance of effect latency or location. Psychophysiology 56, e13335. doi:10.1111/psyp.13335.

Schönbrodt, F.D., Perugini, M., 2013. At what sample size do correlations stabilize? J. Res. Personal. 47, 609–612. doi:10.1016/j.jrp.2013.05.009.

Schönbrodt, F.D., Wagenmakers, E.J., 2018. Bayes factor design analysis: planning for compelling evidence. Psychon. Bull. Rev. 25, 128–142. doi:10.3758/s13423-017-1230-y.

Schönbrodt, F.D., Wagenmakers, E.J., Zehetleitner, M., Perugini, M., 2017. Sequential hypothesis testing with Bayes factors: efficiently testing mean differences. Psychol. Methods 22, 322–339. doi:10.1037/met0000061.

Smith, S.M., Nichols, T.E., 2009. Threshold-free cluster enhancement: addressing problems of smoothing, threshold dependence and localisation in cluster inference. Neuroimage 44, 83–98. doi:10.1016/j.neuroimage.2008.03.061.

Snapinn, S., Chen, M.G., Jiang, Q., Koutsoukos, T., 2006. Assessment of futility in clinical trials. Pharm. Stat. 5, 273–281. doi:10.1002/pst.216.

Spiegelhalter, D.J., Freedman, L.S., Blackburn, P.R., 1986. Monitoring clinical trials: conditional or predictive power? Control. Clin. Trials 7, 8–17. doi:10.1016/0197-2456(86)90003-6.

Tang, D.I., Geller, N.L., 1999. Closed testing procedures for group sequential clinical trials with multiple endpoints. Biometrics 55, 1188–1192. doi:10.1111/j.0006-341X.1999.01188.x.

Van Essen, D.C., Smith, S.M., Barch, D.M., Behrens, T.E.J., Yacoub, E., Ugurbil, K., 2013. The WU-Minn human connectome project: an overview. NeuroImage, Mapping the Connectome 80, 62–79. doi:10.1016/j.neuroimage.2013.05.041.

Wald, A., 1992. Sequential tests of statistical hypotheses. In: Kotz, S., Johnson, N.L. (Eds.), Breakthroughs in Statistics: Foundations and Basic Theory. Springer Series in Statistics. Springer, New York, NY, pp. 256–298. doi:10.1007/978-1-4612-0919-5_18.

Wassmer, G., Pahlke, F., 2020. rpact: confirmatory adaptive clinical trial design and analysis.

Yarkoni, T., Poldrack, R.A., Nichols, T.E., Van Essen, D.C., Wager, T.D., 2011. Large-scale automated synthesis of human functional neuroimaging data. Nat. Methods 8, 665–670. doi:10.1038/nmeth.1635.

Zalesky, A., Fornito, A., Bullmore, E.T., 2010. Network-based statistic: identifying differences in brain networks. Neuroimage 53, 1197–1207. doi:10.1016/j.neuroimage.2010.06.041.