Four Neuroimaging Questions that P-Values Cannot Answer (and Bayesian Analysis Can)

Maxime Taquet, Jurriaan M. Peters, Simon K. Warfield

Computational Radiology Laboratory, Boston Children's Hospital, Harvard Medical School

Abstract. Null Hypothesis Significance Testing (NHST) is used pervasively in neuroimaging studies, despite its known limitations. Recent critiques to these tests have mostly focused on technical issues with multiple comparisons and difficulties in interpreting p-values. While these critiques are valuable, we believe that they overlook the fundamental flaws of NHST in answering research questions. In this paper, we review major limitations inherent to NHST that we formulate as four research questions insoluble with p-values. We demonstrate how, in theory, Bayesian approaches can provide answers to such questions. We discuss the implications of these questions as well as the practicalities of such approaches in neuroimaging.

Keywords: Bayesian, p-value, NHST, Type M, Type S, Prior

1 Introduction

The finding that statistically significant fMRI signal change can be mistakingly observed in a dead salmon performing a mentalizing task [1] and the account of too-high-to-be-true correlations between self-reported behavioral measures and brain activations [19] sparked an heated debate in the brain imaging community about the statistical practice employed in such studies [12,15]. Up to very few exceptions (e.g., [13]), this debate has focused on publication bias, appropriate corrections for multiple comparisons, and reporting of findings in good faith, thereby joining the broader discussion on flawed scientific standards [8]. While defining and advocating good scientific practice is of paramount importance, we feel that this discussion has often diverted the attention from the inappropriateness of null-hypothesis significance testing (NHST) in answering research questions —no matter how meticulously applied. There are indeed important neuroimaging research questions to which NHST provides misleading, if any, answers. In this paper, we review four such questions and we describe how Bayesian methods alleviate the fallacies of NHST. Section 2 recalls the rationale behind NHST and p-values. Section 3-6 review four questions insoluble with p-values. Section 7 discusses the implications of these questions for the neuroimaging community.

2 Brain imaging is free of Type I and Type II errors

NHST proceeds as a proof by contradiction. If our data are incompatible with some hypothesis, then the hypothesis must be wrong. In empirical science, such as neuroimaging, a definite claim of incompatibility cannot be achieved and we therefore compute the probability to observe the data (or something even less compatible with the hypothesis), should the hypothesis be true. This probability is called the *p-value* and the hypothesis is called the *null hypothesis*. If the p-value is small enough, then the null hypothesis is rejected with confidence. If, notwithstanding the small p-value, the null hypothesis was actually true, then one makes a Type I error (rejecting a true null hypothesis).

Imagine that we are interested in the interaction between cinephilia (a passionate interest in cinema) and hippocampal volume. We define a null hypothesis (that cinephiles have on average exactly the same hippocampal volume as control subjects), collect MRI data, compute the tissue volume and attempt to refute the null hypothesis. Now, before we endeavor to do so, we may ponder the odds that the null hypothesis is actually true. This probability most likely equals 0%. Cinephiles tend to enjoy more esoteric movies typically played in smaller theaters for which signs are not displayed in the city. This may require cinephiles to develop better spatial navigation skills, which are associated with larger hippocampi [14]. Whether this line of reasoning prevails in the global association between cinephilia and the volume of hippocampi or other opposite effects play a more important role, the probability that there is absolutely no effect of cinephilia on hippocampal volume is essentially null. The upside of this fact is that most researchers in neuroimaging make no Type I errors nor Type II errors (since null hypotheses are always wrong). The downside of it, however, is that the conclusions of NHST in this context are fairly useless, since the null hypothesis can be rejected prior to acquiring any data. Making no Type I nor Type II errors in brain imaging does not imply that we do not make any error. Our errors pertain to the sign and the magnitude of our conclusions, coined Type S and Type M errors by Gelman *et al.* [5,6].

3 Type S Errors: How confident are we that our finding is not opposite to the truth?

Let us assume that we want to compare the brain connectivity between patients with autism and controls. After comparing the groups, we find out that patients with autism have, on average, significantly weaker connections in the language system ($p < 0.05$). Given this statistically significant result, what are the chances that patients with autism actually have stronger connections in the language system (*i.e.*, what are the odds that my finding is opposite to the truth)? The answer is "we really don't know". To understand why, let us formalize the problem[1]. Let

[1] This formalization is greatly inspired from the formalism in [5] that we adapt to better reflect the situation of image-based population studies of the brain, in which more information is available a priori for control subjects than for patients.

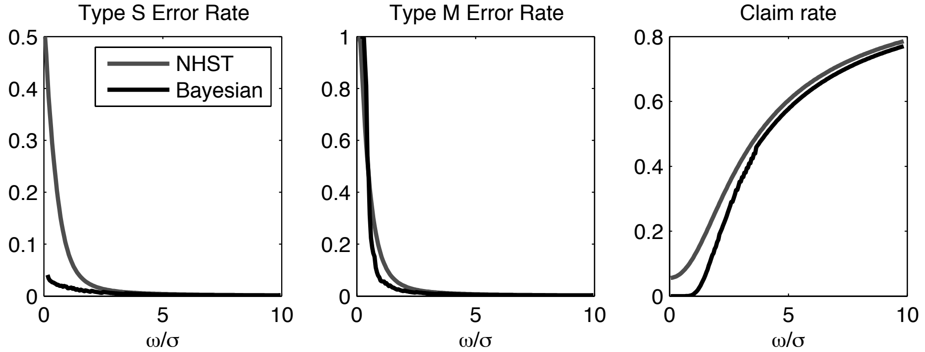

Fig. 1. (Left) NHST makes up to 40% Type S errors for small values of the ratio ω/σ. By contrast, Bayesian analysis controls the Type S error rate to remain below 2.5%. (Middle) Both NHST and Bayesian approaches can make up to 100% Type M errors for small values of ω/σ. (Right) However, for such small values of the ratio, Bayesian approaches are much more prudent than NHST, making very few claims with confidence and showing adaptability to the data as ω/σ increases.

θ_1 and θ_2 be the true mean connectivity in the language system of patients with autism and controls respectively. We assume that $\theta_1, \theta_2 \in [-\infty, \infty]$. Let y_1 and y_2 be the observed mean connectivity in patients with autism and in controls. Assuming normality and equal variance (σ^2) in both groups, we have:

$$y_i|\theta_i, \sigma \sim \mathcal{N}(\theta_i, \sigma^2), \quad i = 1, 2. \tag{1}$$

Type S errors occur whenever $y_1 - y_2$ reaches a specific threshold T to make a claim with confidence (e.g., $|y_1 - y_2| > 1.96\sqrt{2}\sigma$ corresponding to $p < 0.05$) while having a sign that is opposite to the true difference $\theta_1 - \theta_2$. The probability of making a Type S error is therefore given by [5]:

$$P\Big(\text{Type S error}\Big) = P\left(\text{sign}(y_1 - y_2) \neq \text{sign}(\theta_1 - \theta_2)\Big| |y_1 - y_2| > T\right). \tag{2}$$

This probability involves computing the posterior probability over the latent variables θ_i and can therefore only be estimated in a Bayesian approach. Now, although this probability cannot be directly estimated in NHST, is the p-value returned by the test a good enough proxy to the Type S error rate? The answer, as we describe below, is "No".

To estimate the probability in (2), let us define a simple hierarchical model as in [5]. We assume that σ can reliably be estimated from the data. The prior on θ_1 and θ_2, $p(\theta_1, \theta_2)$, can be expressed conditionally: $p(\theta_1)p(\theta_2|\theta_1)$. The prior $p(\theta_1)$ encodes any prior knowledge that we have about the mean connectivity in controls, as gained, for example, from past experience with such connectivity measures. We may, for instance, assign a normal prior to θ_1 centered at some reasonable μ and some standard deviation τ (our conclusions, as we will see, depend neither on the value μ nor on that of τ):

$$\theta_1|\mu, \tau \sim \mathcal{N}(\mu, \tau^2). \tag{3}$$

If we knew the true value of the mean connectivity in controls, θ_1, and we had no data from patients, our best guess about the value in patients, θ_2, would be θ_1. We can therefore model the conditional prior on θ_2 as a distribution centered on θ_1, for example a normal with unknown variance ω^2 (we will come back to estimations of the value of ω):

$$\theta_2|\theta_1, \omega \sim \mathcal{N}(\theta_1, \omega^2). \tag{4}$$

From the hierarchical model described by (1),(3) and (4), we can derive the posterior probability of $\delta \triangleq \theta_1 - \theta_2$ given $d \triangleq y_1 - y_2$ and the joint probability of δ and d:

$$\delta|d, \omega, \sigma \sim \mathcal{N}\left(\frac{d}{1+\frac{2\sigma^2}{\omega^2}}, \frac{1}{\frac{1}{2\sigma^2}+\frac{1}{\omega^2}}\right) \tag{5}$$

$$[d, \delta]|\omega, \sigma \sim \mathcal{N}\left(\mathbf{0}, \omega^2\begin{pmatrix}1+\frac{2\sigma^2}{\omega^2} & 1\\ 1 & 1\end{pmatrix}\right). \tag{6}$$

Equation (5) allows us to define a 95% posterior interval on δ and therefore define a threshold T_B to make a claim with confidence about the sign of the difference:

$$T_B = 1.96\sqrt{2}\sigma\sqrt{1+\frac{2\sigma^2}{\omega^2}}. \tag{7}$$

Equation (6) then allows us to estimate the Type S error rate for a given threshold T by conditioning on the fact that $|d| > T$ [5]:

$$P\Big(\text{Type S error}\Big) = \frac{\displaystyle\int_{-\infty}^{0}\int_{T}^{\infty} p([d,\delta]|\omega,\sigma)dd\,d\delta + \int_{0}^{\infty}\int_{-\infty}^{-T} p([d,\delta]|\omega,\sigma)dd\,d\delta}{\displaystyle\int_{-\infty}^{+\infty}\int_{T}^{\infty} p([d,\delta]|\omega,\sigma)dd\,d\delta + \int_{-\infty}^{\infty}\int_{-\infty}^{-T} p([d,\delta]|\omega,\sigma)dd\,d\delta}.$$

The conditioning on $|d| > T$ means that we consider a Type S error only if we make an incorrect claim with confidence. This error rate only depends on ω/σ and is depicted in Fig. 1 for $T_F = 1.96\sqrt{2}\sigma$ corresponding to $p < 0.05$ in NHST and for T_B given in Equation (7). For values of $\omega \gg \sigma$, $T_B \approx T_F$ so that both inferences lead to similar Type S errors. However, for $\omega \lesssim \sigma$, the Type S error rate with NHST grows quickly and reaches 40% for $\omega = 0.15\sigma$, demonstrating that we really don't know the odds of making a Type S error given some p-value. By contrast, Type S error rates with the Bayesian approach remain below 5% for all values of ω and is therefore under control.

The ratio ω/σ encodes how far apart we expect, a priori, the variables θ_1 and θ_2 to be with respect to the variance of observations y_1 and y_2. In other words, this ratio encodes our prior on the true underlying effect size and we have:

$$E\left[\left(\frac{\delta}{\sigma}\right)^2\right] = \left(\frac{\omega}{\sigma}\right)^2.$$

NHST can therefore be interpreted as a Bayesian approach which assumes that, a priori, effect sizes are infinite on average. This assumption seems unreasonable and would lead to the observation of extreme group differences in almost all cases. This explains why the actual Bayesian approach performs better for all finite values of ω/σ (and equivalently to NHST for infinite values of ω/σ) as shown on Fig. 1. The actual value of ω could be estimated by pooling all comparisons (all connections) made between the brain of controls and patients. This first example illustrates that NHST cannot resolve important aspects of inference in population studies whereas Bayesian inference enables more adaptive and reliable analyses of the data at hand.

4 Type M Errors: Can the true effect be much smaller than what we observed?

Suppose that we are confident (for some ad-hoc reason) that our finding is not a Type S error. Since we never make any Type I and Type II error in brain imaging, what else can invalidate our finding? We may have observed too strong an effect compared to the true effect. This would be a Type M error [6]: the sign of the observed effect may be correct but its magnitude is not.

Again, the question of the prevalence of such errors in practice cannot be answered using NHST alone but can be answered in a Bayesian fashion. Using the same hierarchal model as in the previous section (Equations (1), (3) and (4)), and the resulting posterior and joint distributions (Equations (5) and (6)), we can estimate the Type M error rate. If we define a Type M error as misestimating the effect by a factor 10 ($|d| < |\delta|/10$ or $|d| > 10|\delta|$) while claiming this effect with confidence (*i.e.*, conditionally on $|d| > T$), then the Type M error rate is given by:

$$P\Big(\text{Type M error}\Big) = \frac{\displaystyle\int_T^\infty \int\limits_{\substack{[-\infty,\frac{d}{10}]\\ \cup[10d,\infty]}} p([d,\delta]|\omega,\sigma)d\delta\,dd + \int_{-\infty}^{-T} \int\limits_{\substack{[-\infty,10d]\\ \cup[\frac{d}{10},\infty]}} p([d,\delta]|\omega,\sigma)d\delta\,dd}{\displaystyle\int_{-\infty}^{+\infty}\int_T^\infty p([d,\delta]|\omega,\sigma)dd\,d\delta + \int_{-\infty}^{\infty}\int_{-\infty}^{-T} p([d,\delta]|\omega,\sigma)dd\,d\delta}.$$

The threshold T used in this equation depends on the inference approach being used. For NHST, the conventional 95% confidence interval gives rise to $T_F = 1.96\sqrt{2}\sigma$, whereas the Bayesian approach leads to a 95% posterior interval governed by the threshold T_B of Equation (7). For these thresholds, the Type M error rates are illustrated in Fig. 1. Interestingly, unlike Type S error rates, Type M error rates reach high levels for both NHST and Bayesian approaches for small values of ω/σ. When this ratio falls under 0.45, Type M error rates for both Bayesian and NHST are above 50%, implying that every other finding has an effect that is at least an order of magnitude off compared to the true effect. Not surprisingly, overestimation errors (that is $|d| > 10|\delta|$) are overwhelmingly more present than underestimations in this range (for $\omega/\sigma < 0.5$, approximately all Type M errors consists of overestimations).

Given that Type M errors can reach dramatically high effects with both NHST and Bayesian approaches, one may question what we have gained from the Bayesian approach. There are three reasons for which the Bayesian approach remains beneficial in this case. First, without a Bayesian approach, we would not have been aware of the prevalence of Type M errors, the estimation of which requires the introduction of a prior over latent variables θ_i. Second, for $\omega/\sigma > 0.5$, Type M error rates are consistently smaller with the Bayesian approach than with NHST. Third, and most importantly, we have defined Type M error rates conditionally on making a claim with confidence. Since the threshold for confidence differs between NHST and the Bayesian approach, so will the number of claims being made with confidence. Fig. 1 depicts the rate of claims. Strikingly, the Bayesian approach makes almost no claim with confidence for small values of ω/σ whereas NHST makes at least 5% in all cases (as a consequence of the definition of the p-value). As the ratio ω/σ increases, the number of claims made with the Bayesian approach increases to become closer to the number of claims made with NHST. In other words, the Bayesian approach is always more conservative than NHST (since $T_B > T_F$) and is even more conservative –and rightly so– when the claims will likely lead to a Type M error. This finding shows that, by zealously controlling for hypothetical Type I errors, NHST makes substantially more actual Type M errors. On the other hand, by properly modeling the uncertainty of observations and priors on effect sizes, Bayesian approaches adopt an adaptive behavior in which fewer claims are made with confidence when the data does not justify them.

5 Do patients and controls have similar brains?

The probability that two groups of individuals have exactly the same average brain is most often zero. That is because statistics with a continuous domain (for example, the hippocampal volume) often have no mass at zero. In other words, this is because the integral between zero and zero of a finite function is zero. Yet, we do not expect the brains of all patients in all diseases to be affected in all its properties and in all its locations. We expect to observe some *similarities* between brains. But what does *similar* mean in a world where null hypotheses are intrinsically impossible? As we shall see, the answer to this question is not so much statistical as it is biomedical.

First, let us recall why large p-values are no evidence that brains are similar despite its occasional use as such in population studies. Take a somehow well-established neurological finding, for example that patients with agenesis of the corpus callosum (AgCC, the complete or partial absence of a corpus callosum)

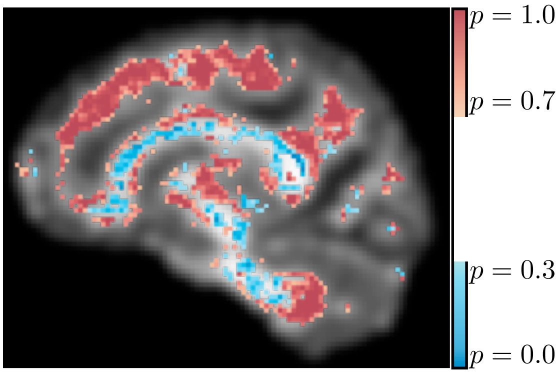

Fig. 2. Example of a map of the posterior probability that the feature of interest (here, the radial diffusivity from [18]) is out of the ROPE: large values are evidence that there is an important difference between the groups (red areas) whereas small values are evidence that the groups are similar (blue areas). The latter cannot be observed in NHST since p-values in those areas are uniformly distributed (and not specially high).

have disrupted functional inter-hemispheric connections. Now, recruit patients with AgCC and healthy controls, disregard their gender, age and ethnicity, acquire fMRI images, align images to an atlas based on rigid registration only and perform subsequent processing using a version of SPM99 in which a grad student of your lab mistakingly introduced some bugs. In that scenario, the odds of getting a p-value larger than 0.05 are approximately 95% despite the fact that there actually is a true substantial difference between the groups. Furthermore, very large p-values (e.g., $p > 0.9$) occur randomly if the difference between groups is very little (under the extreme case of zero effect, p-values are uniformly distributed). We therefore cannot increase the threshold on p in a hope to better detect similarities: setting a threshold at $p > 0.95$ would result in at least 95% missed detections of similarities.

If similarities do not imply zero effect, what do they imply? We submit that brains are similar if the difference between them falls within a region of practical equivalence (ROPE) [10] as defined in a Bayesian context in [11]. We therefore want to estimate the probability that the group difference $\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2$ is within the ROPE:

$$P((\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2) \in \text{ROPE}|\text{Data})$$

This probability is computed from the posterior distribution over the latent variables and is therefore only computed in a Bayesian framework. Large values of this posterior are evidence that brains are similar between the groups up-to a difference that is within the ROPE (Fig. 2). The definition of the ROPE is not, however, informed by any statistical argument and should rather be answered in a biomedical context. We propose two avenues to define the ROPE: a literature-based based and an introspection-based approach.

Literature-based ROPE We can establish the ROPE based on other existing studies. For instance, if we want to compare the volume of the hippocampus of patients suffering Alzheimer's disease, we may want to consider that, training for a year as a taxi driver already changes the hippocampal volume by approximately 1% [14]. Since we expect more dramatic changes to occur in the brain of patients with Alzheimer's disease than in the brain of taxi drivers, we may consider brains that differ in their hippocampal volume by up to 1% to be practically equivalent. For studies in pediatrics, one particularly interesting condition to establish relevant orders of magnitude is *age*. If we have curves for the evolution of different brain properties as children develop, we may consider a difference corresponding to one week, one month or one year of maturation as being within the ROPE.

Introspection-based ROPE In the absence of previous relevant studies, one may ponder the embarrassment involved should the actual magnitude of a statistically significant difference be reported in addition to the *p*-value. In neuroimaging, *p*-values are sometimes reported without the actual magnitude of the effect. Imagine that, while measuring the volume of the hippocampus, we observe a difference between controls and patients that is tiny (e.g., 0.06mm^3) yet statistically significant. Would we feel comfortable reporting such a tiny effect? If the answer is "No", then we must probably consider such a difference as belonging to the ROPE. This strategy was used in [18] to set ROPE to microstructural differences in the brain (difference in fascicle directions, diffusivities and volumetric fractions).

6 What is the probability that the patient has the disease?

One goal of the definition of biomarkers through population studies is to assist physicians in the decision-making process. In this process, brain images only constitute part of the available information. When making a diagnosis, physicians start off with such information as the patient's age, gender, history and clinical assessment. For instance, the same observed abnormality of a patient's hippocampus is not as strong a sign of Alzheimer's disease in a patient who is 55 years old as it is in a patient who is 68 years old. The odds for the patient to have the disorder also increase if the hippocampal abnormality co-occurs with a clinical presentation of memory impairment or with the presence of a first-degree relative with the disorder. How can all this information be used to estimate the probability that the patient has the disease?

NHST would proceed by eliminating all possible null hypotheses (the null hypothesis that the patient has no disease, then all other null hypotheses that the patient has any other kinds of dementia). After contradicting all null hypotheses, we may believe that Alzheimer's disease is the only possible hypothesis that holds, and yet we would not have any idea of the probability of it being true.

This is akin to a diagnosis of exclusion, used in medical practice when no direct conclusive diagnosis can be made.

In the Bayesian formalism, the probability of the disease (D) given all pieces of information can be computed as a posterior probability from the likelihood of the brain imaging data (B) and the prior of the disease given clinical (C) and other individual data (I):

$$\begin{aligned} P(D|B,C,I) &= \frac{P(B|D,C,I)P(D|C,I)}{P(B|C,I)} \\ &= \frac{P(B|D,C,I)P(C|D,I)P(D|I)}{P(B|C,I)P(C|I)} \\ &\propto P(B|D,C,I)P(C|D,I)P(D|I) \end{aligned} \tag{8}$$

Since the denominator does not depend on D and since D is a binary variable, it is sufficient to compute the numerator for both $D = 1$ (has the disease) and $D = 0$ (does not have the disease) and to infer the denominator from the fact that $P(D = 1|B,C,I) + P(D = 0|B,C,I) = 1$. The factor $P(B|D,C,I)$ can be inferred from a model of some kind (for example a generalized linear model), $P(C|D,I)$ can be inferred by calibrating the clinical assessments protocols (for example, those of DSM-V) and $P(D|I)$ is the prevalence of the disease and can usually be obtained from large public health surveys.

Equation (8) stands as a theoretical framework to infer the probability of a patient having the disease given her brain images, her personal and historical information and her clinical assessment. This framework relies on a hierarchical Bayesian model in which prior information can naturally be integrated.

7 Discussion

Throughout the last four sections, we demonstrated that p-values are not appropriate to answer some important brain imaging questions. They fail to predict Type S and Type M error rates (which depend on the prior over effect sizes that is assumed infinite in NHST), they fail to provide evidence for similarities between brains and they cannot be used to estimate the probability that a patient has a disease. We have shown how Bayesian hierarchical models naturally answer those questions. In this section, we discuss the practical implications of these considerations for the neuroimaging research community.

What does Bayesian analysis tell us about the appropriate sample size? Intuitively, most researchers would probably agree that the more data we have the better the inference. Yet this idea was challenged by Friston in his *Ten Ironic Rules for Non-Statistical Reviewers* [2] on the ground that, for a constant p-value, a significant finding based on fewer samples implies a larger effect size. In t-tests for instance, the p-value is a strictly decreasing function of $d/\sigma = \sqrt{n}d/\sigma'$ where $\sigma' = \sqrt{n}\sigma$ is the standard deviation of the individual samples whereas σ is the standard deviation of their mean in each group (denoted y_1 and y_2 in

Section 3 and 4). Increasing n while keeping the p-value constant thus implies decreasing d and therefore decreasing the effect size d/σ'.

However, lower n implies higher σ (since $\sigma = \sqrt{n}\sigma'$ and σ' is determined by the measurement noise and inter-subject variability) which implies lower ω/σ ratios. At lower ω/σ ratios, the Type S and Type M error rates are dramatically higher so that the inference is less reliable as described in Sections 3 and 4. What Friston describes as *larger effect sizes* should be understood as *larger observed effect sizes* which may actually correspond to a *smaller true effect size* likely affected by a Type M error (or even a Type S error), an effect known as p-value filter bias [6] or inflated early-effect sizes [9].

Acknowledging this fallacy of classical inference, Friston further argues that there ought to be a compromise in the choice of a sample size [2]. Too small a sample size would likely lead to inflated early effect-sizes, whereas too high a sample size would result in the detection of trivial effects (*i.e.*, statistically significant effects that are too small to be interesting). Such effects can naturally be accounted for in a Bayesian framework by considering them as practically equivalent to no effect (as described in Section 5). The more data we acquire, the more confident we are that the effect is within or out-of the ROPE and the more accurate our estimate of its sign (fewer Type S errors) and its magnitude (fewer Type M errors). Our Bayesian account of this question therefore indicates that indeed the more data we have, the better.

Bayesian or frequentist inference? This paper should not be understood as a critique of frequentist inference as a whole. We rather question the appropriateness of NHST in a non-dichotomous context such as brain imaging, much like other researchers have questioned it before in other contexts such as political science [3]. When zero-effects never actually occur, we believe that there is no reason to try hard to control for Type I error rates. Our disagreement with the rationale of NHST can equally be expressed for Bayesian dichotomous analyses such as Bayesian $t-$tests [16] that provide mechanisms to "accept or reject the null hypothesis". In contrast, we believe that there may be some purely frequentist approaches (such as a classifier learned and validated by bootstrapping) that may be appropriate to draw interesting conclusions from brain imaging data, including, for example, to answer the question in Section 6.

Bayesian analysis of neuroimaging data. We believe that the natural answers brought by Bayesian approaches to the four research questions presented above should encourage the neuroimaging community to develop novel Bayesian inference methods. The development of such methods comes with their share of hurdles to overcome. These difficulties pertain to the need for a balance between accurate representation of the data and computational tractability [4]. The challenges in representing brain imaging data as a tractable hierarchical model arises from the *within-voxel* complexity of variables and *between-voxel* dependencies between them.

The increasing complexity of the information contained *within* each voxel leads variables that often belong to non-trivial spaces and with non-trivial dependencies between them. For example, in microstructure imaging, each voxel contains a complete model of the brain microstructure that may present with ten or more variables. These variables belong to the sphere (for directions), the space of strictly positive numbers (for diffusivities and dispersion coefficients) and the simplex (for signal fractions) [18]. A Bayesian hierarchical model should account for these particular spaces. Other examples include the time series contained in fMRI voxels in which contiguous time points are dependent in a non-trivial manner.

Brain imaging variables are also statistically dependent *between* neighboring voxels. In theory, this dependency can be captured by a graphical model, such as a discrete Markov random field [20]. However, graphical models substantially increase the computational complexity when estimating posterior probabilities since the inference needs to be done globally instead of voxel-wise. Typically, in those cases, approximations such as Variational Bayes (VB) methods are employed instead of sampling strategies such as Markov Chain Monte Carlo (MCMC) [20]. These approximations have been shown to outperform non-Bayesian introductions of spatial information in problems such as image registration [17]. However, they are known to introduce biases in the estimations of posterior probabilities [7]. These biases may be of little concern when the inference results of interest are specific values of some variables (e.g., the prior components and the deformation field in [17]) because these values may be exact even when the posterior probability estimate is not. Biases in estimates of the posterior may be more concerning when the goal of inference is to obtain the actual value of the posterior probability, as when assessing the probability that a patient has some disease. In those cases, MCMC sampling and its computational burden may be unavoidable or the bias caused by VB methods should be proved negligible. We believe that these are important avenues for future research.

8 Conclusion

Null hypothesis significance testing (NHST) is a well-defined concept that, if properly conducted, results in a mathematically correct p-value. This p-value, in neuroimaging, is however often useless since all null hypotheses can readily be refuted on the basis that any condition affects our brain in some way. There are therefore many important research questions in neuroimaging that NHST cannot answer. This paper has reviewed these questions and proposed Bayesian alternatives to answer them. Bayesian approaches reduce inference errors (Type S and Type M), enable the building of evidence for the presence of similarities between brains and the incorporation of prior information in the diagnosis, as gleaned from clinical history and examinations. To leverage Bayesian approaches in neuroimaging analyses, technical difficulties related to the complexity of the information within and between voxels must be overcome. Important methodological developments in this area are being made and should be sustained and expanded to move the neuroimaging community away from the inappropriateness of NHST.

References

1. Bennett, C.M., Miller, M., Wolford, G.: Neural correlates of interspecies perspective taking in the post-mortem atlantic salmon: An argument for multiple comparisons correction. In: Organization for Human Brain Mapping. pp. S39–S41 (2009)
2. Friston, K.: Ten ironic rules for non-statistical reviewers. Neuroimage 61(4), 1300–1310 (2012)
3. Gelman, A.: Commentary: p values and statistical practice. Epidemiology 24(1), 69–72 (2013)
4. Gelman, A., Shalizi, C.R.: Philosophy and the practice of bayesian statistics. British Journal of Mathematical and Statistical Psychology 66(1), 8–38 (2013)
5. Gelman, A., Tuerlinckx, F.: Type S error rates for classical and bayesian single and multiple comparison procedures. Computational Statistics 15(3), 373–390 (2000)
6. Gelman, A., Weakliem, D.: Of beauty, sex and power. American Scientist 97(4), 310–316 (2009)
7. Hoffman, M.D., Gelman, A.: The no-u-turn sampler: Adaptively setting path lengths in hamiltonian monte carlo. arXiv preprint arXiv:1111.4246 (2011)
8. Ioannidis, J.P.: Why most published research findings are false. PLoS medicine 2(8), e124 (2005)
9. Ioannidis, J.P.: Why most discovered true associations are inflated. Epidemiology 19(5), 640–648 (2008)
10. Jones, B., Jarvis, P., Lewis, J., Ebbutt, A.: Trials to assess equivalence: the importance of rigorous methods. Bmj 313(7048), 36–39 (1996)
11. Kruschke, J.K.: Bayesian assessment of null values via parameter estimation and model comparison. Perspectives on Psychological Science 6(3), 299–312 (2011)
12. Lieberman, M.D., Berkman, E.T., Wager, T.D.: Correlations in social neuroscience aren't voodoo: commentary on Vul et al. Perspectives on Psychological Science 4(3), 299–307 (2009)
13. Lindquist, M.A., Gelman, A.: Correlations and multiple comparisons in functional imaging: a statistical perspective (commentary on vul et al., 2009). Perspectives on Psychological Science 4(3), 310–313 (2009)
14. Maguire, E.A., et al.: Navigation-related structural change in the hippocampi of taxi drivers. Proc. Nat. Acad. Sci. 97(8), 4398–4403 (2000)
15. Nichols, T.E.: Multiple testing corrections, nonparametric methods, and random field theory. Neuroimage 62(2), 811–815 (2012)
16. Rouder, J.N., Speckman, P.L., Sun, D., Morey, R.D., Iverson, G.: Bayesian t tests for accepting and rejecting the null hypothesis. Psychonomic bulletin & review 16(2), 225–237 (2009)
17. Simpson, I.J., Woolrich, M.W., Cardoso, M.J., Cash, D.M., Modat, M., Schnabel, J.A., Ourselin, S.: A bayesian approach for spatially adaptive regularisation in non-rigid registration. In: Medical Image Computing and Computer-Assisted Intervention–MICCAI 2013, pp. 10–18. Springer (2013)
18. Taquet, M., Scherrer, B., Peters, J.M., Prabhu, S.P., Warfield, S.K.: A fully bayesian inference framework for population studies of the brain microstructure. In: Medical Image Computing and Computer-Assisted Intervention–MICCAI 2014. Springer (2014)
19. Vul, E., Harris, C., Winkielman, P., Pashler, H.: Puzzlingly high correlations in fmri studies of emotion, personality, and social cognition. Perspectives on psychological science 4(3), 274–290 (2009)
20. Woolrich, M.W., et al.: Bayesian analysis of neuroimaging data in FSL. Neuroimage 45(1), S173–S186 (2009)