# A hybrid procedure for detecting Global Treatment Effects in Multivariate Clinical Trials
# Theory and Applications to fMRI Studies

Giorgos Minas[1,2], Fabio Rigat[1,2,3], Tom Nichols[4], John Aston[1], Nigel Stallard[3]

[1] Department of Statistics, University of Warwick, Coventry, UK
[2] Warwick Centre of Analytical Sciences, University of Warwick, Coventry, UK
[3] Novartis Vaccines and Diagnostics, Siena, Italy
[4] Health Sciences Research Institute, Warwick Medical School, University of Warwick, Coventry, UK
[5] Neuroimaging Statistics, Warwick Manufacturing Group & Department of Statistics, University of Warwick, Coventry, UK

June 2010

**Abstract**

In multivariate clinical trials, a key research endpoint is ascertaining whether a candidate treatment is more efficacious than an established alternative. This global endpoint is clearly of high practical value for studies, such as those arising from neuroimaging, where not only the outcome dimensions are numerous but also highly correlated and the available sample sizes are typically small. In this paper, we develop a two-stage procedure testing the null hypothesis of global equivalence between treatments effects and we demonstrate its application to analysing phase II neuroimaging trials. Prior information such as suitable statistics of historical data or suitably elicited expert clinical opinions are combined with data collected from the first stage of the trial to learn a set of optimal weights. These weights are applied to the outcome dimensions of the second-stage responses to form the linear combination $z$ and $t$ test statistics while controlling the test's false positive rate. We show that the proposed tests hold desirable asymptotic properties and we characterise their power functions under wide conditions. In particular, by comparing the power of the proposed tests with that of Hotelling's $T^2$ we demonstrate their advantages when sample sizes are close to the dimension of the multivariate outcome. We apply our methods to fMRI studies where we find that, for sufficiently precise first stage estimates of the treatment effect, standard single-stage testing procedures are outperformed.

**keywords:**clinical trials, multivariate outcomes, hypothesis testing, hybrid methods, fMRI

## 1 Introduction

The development of high throughput biological assays lies at the heart of the emerging concept of personalised medicine [29], [10]. Two main areas where high throughput technology is rapidly improving detection accuracy are gene expression profiling [14], [43] and medical imaging [1], [49], [15], [32]. The clinical application of these techniques is one of the main drivers of major changes in drug discovery and development [5], [30] as well as the identification of surrogate endpoints in clinical trials [37].

In an effort to support these experimental and technological developments, in this paper we focus on developing statistical methodology for detecting global treatment effects from high throughput functional magnetic resonance imaging (fMRI) data. Clinical applications of fMRI so far include Alzheimer disease [3], schizophrenia [16], depression [7], addiction [4] and pain treatment [21]. More widespread clinical application of fMRI is currently hampered by a partial understanding of the relation between fMRI data and neuronal activity and by a relatively low signal-to-noise ratio [23]. Our motivations are eminently practical, in that the detection of global effects is one of the main endpoints in drug development [45], [22] and the fMRI signal is being considered as a potential surrogate endpoint in phase II proof of concept clinical trials [55], [17], [54].

This paper is organised as follows. In Section 2 we briefly describe fMRI signal and illustrate the advantages and limitations of standard testing procedures for detecting global treatment effects using fMRI Regions of Interest (ROI) data. In Section 3 we develop a novel testing procedure maximizing the power of detecting global treatment effects under more general conditions with respect to current multivariate approaches. The proposed procedure is based on a linear combination of the ROI multivariate responses where the weights of each ROI

are estimated from a pilot study and the test statistic is calculated using data arising from a subsequent main study. The main properties of the power function of our procedure are explored in section 4. We show that the optimal proportion of the total samples allocated to weight estimation can lie well below 50%, allowing for accurate testing even for phase I and II clinical trials. In section 4 we also use simulated data generated under various model parameters to compare our procedure with other standard tests, including the $z$, $t$ and Hotelling's $T^2$ statistics. Here we show that our procedure can achieve high detection power even when the sample size approaches the response dimension, that is when Hotelling's $T^2$ either lacks power or is not applicable. Section 4 is closed with the analysis of real fMRI recordings, showing that, in this example, the proposed procedure attains substantially lower p-values than the alternative tests. Section 5 concludes this paper with a critical discussion of its main theoretical and empirical results and with selected directions for future work.

## 2 Single stage analysis of fMRI clinical trials

Blood oxygenation level dependent (BOLD) fMRI records brain activity by detecting the accompanying changes in the local level of blood oxygenation. fMRI signal is measured at a spatial resolution of 2-4 mm and on a 1-4 s temporal resolution allowing for a study of brain systems at biologically meaningful spatio-temporal scales [19]. The raw fMRI data are preprocessed using techniques of signal processing, image processing and statistics. This typically involves eliminating the most common experimental artifacts (e.g. motion and scanner artifacts), normalisation (e.g. using spatial smoothing) and registration of the raw data at a common reference image [41]. The preprocessed data consist of series of 3-dimensional brain images recording the BOLD contrast at voxel[1]-by-voxel resolution. In neurological clinical trials the preprocessed fMRI data are often used to inform a Regions of Interest (ROI) analysis, where the voxel-specific treatment effect estimates, extracted from mass univariate models (see figure 1), are averaged over relatively large functionally meaningful brain regions [39]. Such a ROI approach produces a manageable number of well-defined outcomes and thus powerful testing procedures can be potentially constructed [33].

A simple method for testing for global treatment effects in ROI analysis is the classical Bonferroni correction [9]. Despite its simplicity, Sankoh et al. [42] and Neuhäuser [34] show that for correlations $\rho \geq 0.5$ the Bonferroni method may become substantially conservative. In these cases the family-wise error rate (FWER), i.e. the probability of falsely rejecting at least one of the local null hypotheses, may be much smaller than its nominal $\alpha$-level [38]. Since high correlations are typically observed [40] when the number of ROI, $K$, grows, the Bonferroni correction may be very inefficient for fMRI data analysis.

A key idea proposed by Westfall et al. [53] is to use reliable prior information for improving the efficiency of multiple testing procedures while controlling the FWER. In this work we follow a similar approach by employing Bayesian inference to construct our test statistic while controlling the false positive rate. In contrast to [53] we incorporate the correlations between ROI into our multivariate modeling assumptions.

The classical approach to univariate testing for multivariate responses is Hotelling's $T^2$ [18], [2]. Hotelling's $T^2$ test is the uniformly most powerful test for multivariate normal responses. However, it is inapplicable when the number of components, $K$, is greater than the sample size, $n$, and it lacks power when $n$ approaches $K$ from above (see section 4). Since this is the typical situation for neurological clinical trials using fMRI, we follow O'Brien's approach in [35] by reducing data dimensionality using a weighted average of the response components with ROI-specific $K$-dimensional weights $w$. Suitable estimates of $w$ result in effective detection of the global treatment effect while inappropriate weights may critically increase false negative rates. In [35] these weights are derived using ordinary least squares (OLS) and generalised least squares (GLS) methods under the assumption of equal size of the treatment effect across the multivariate response. The linear combinations of the components of the response with weighting vectors $w^{OLS} = 1_K = (1, 1, ..., 1)'$ and $w^{GLS} = \Sigma^{-1}1_K$ are used to construct the classical $z$ and $t$ tests [6], [52], [28], [8].

O'Brien's approach to selecting $w$ "reflects the investigator's knowledge about the direction in which the alternative lies" [51]. Thus if a treatment targets mainly one ROI, say the hippocampus for Alzheimer's disease, a suitable weighting vector is the all-zero-but-1 $e_1 = (1, 0, ..., 0)$ where the first component is assigned to the corresponding region. Along the same lines, the weighting vector can be constrained to lie within the space of all-zero-but-$K'$ ($K' < K$) solutions. These solutions are effective in detecting global treatments if the assumed direction of the treatment effect is correct. When this prior belief is incorrect, the false negative rates can be high even in cases where the treatment effect is very large. Alternatively, Lauter et al. [26] propose selecting $w$ using the observed data in such a way that the false positive rate is controlled. The remainder of this paper illustrates a generalization of the tests in [35] and [26] achieving high detection power for arbitrary mean and covariance structures.

[1] voxel: a three-dimensional volume element.

# 3 Testing the global null hypothesis using a 2-stage procedure

Let the $K$-dimensional response vectors $Y_i$ for subjects $i = 1, 2, ..., n_y$ be conditionally independent Gaussian random variables

$$Y_i \overset{\text{iid}}{\sim} \mathcal{N}_K(\mu, \Sigma), \tag{3.0.1}$$

with mean and covariance matrix $\mu = (\mu_1, ..., \mu_K)'$ and $\Sigma = (\sigma_{ik})_{i,k=1}^K$. Gaussianity is an accepted assumption when modeling linear contrasts of normalized fMRI data for ROI analysis [12]. We reduce the dimension of the multivariate response using the scalar linear combination

$$L_i = w'Y_i = \sum_{k=1}^K w_k Y_{ik} \ , \ i = 1, 2, ..., n_y, \tag{3.0.2}$$

where the weighting vector is formally defined as $w = (w_1, w_2, ..., w_K)'$, $w \in \mathbb{R}^K \setminus \{\underline{0}\}$. We wish to test the global null hypothesis of no treatment effect against the two-sided alternative

$$H_0 : \mu = \underline{0} \quad \text{versus} \quad H_1 : \mu \neq \underline{0}. \tag{3.0.3}$$

The $z$- and $t$-statistics for testing $H_0$ against $H_1$ using a random sample $L_i$, $i = 1, ..., n_y$ of the linear combination $L$ respectively when $\Sigma$ is either known or unknown are

$$Z_w = \frac{\overline{L}}{\sigma_L / \sqrt{n_y}}, \tag{3.0.4}$$

and

$$T_w = \frac{\overline{L}}{s_L / \sqrt{n_y}}. \tag{3.0.5}$$

Here, $\sigma_L^2 = w'\Sigma w$, $\overline{L} = \frac{1}{n_y}\sum_{i=1}^{n_y} L_i = w'\overline{y}$ and $s_L^2 = \frac{1}{n_y - 1}\sum_{i=1}^{n_y}\left(L_i - \overline{L}\right)^2 = w'S_y w$ are the variance, sample mean and sample variance of the linear combination $L$, respectively, and $\overline{y} = \frac{1}{n_y}\sum_{i=1}^{n_y} y_i$, $S_y = \frac{1}{n_y - 1}\sum_{i=1}^{n_y}(y_i - \overline{y})(y_i - \overline{y})'$ are the sample mean and covariance matrix of the response $Y$. Under these modeling assumptions, the decision rules

$$\text{reject } H_0 \quad \Leftrightarrow \quad |Z_w| > z_{\alpha/2}, \tag{3.0.6}$$

$$\text{reject } H_0 \quad \Leftrightarrow \quad |T_w| > t_{n_y - 1, \alpha/2}, \tag{3.0.7}$$

specify hypothesis tests of size $\alpha$ for (3.0.3) where $z_{\alpha/2}$ and $t_{n_y-1,\alpha/2}$ are respectively the $100(1-\alpha/2)$ percentiles of the standard Normal and of the Student's t-distribution with $n_y - 1$ degrees of freedom. The power of these tests, i.e. the probability to reject the null hypothesis for given values $\mu$ and $\Sigma$ of the unknown model parameters, are respectively

$$\beta_z(w, \mu, n_y) = Pr_\mu\left(|Z_w| > z_{\alpha/2}\right), \tag{3.0.8}$$

$$\beta_t(w, \mu, \Sigma, n_y) = Pr_{\mu,\Sigma}\left(|T_w| > t_{n_y-1,\alpha/2}\right). \tag{3.0.9}$$

Note that, unlike for Hotelling's $T^2$ test, the degrees of freedom of the linear combination $t$-statistic do not depend on the dimension of the response $K$ and hence the linear combination $t$-test is applicable even for $n_y < K$. Moreover, when $w$ is fixed, both statistics are scale invariant with respect to $w$, i.e. $Z_w = Z_{cw}, T_w = T_{cw}$ for any $c \neq 0$. Also, the distributions of both test statistics are invariant to scale transformations $Y \to cY$ $(c \neq 0)$ of the response. In what follows we focus on estimating the optimal weighting vector maximising the power of the herein described linear combination $z$ and $t$ tests.

## 3.1 The optimal weighting vector

Intuitively, an optimal approach is to select the weighting vector $w$ so as to maximise the probability of detecting the global treatment effects. This is equivalent to selecting $w$ so that the probability of correctly rejecting the global null hypothesis, i.e. the power, is maximised. Theorem 1 describes the form of such an "optimal" weighting vector for fixed values of the unknown parameters without relying on any assumed mean or covariance structure.

**Theorem 1.** *Under* (3.0.1) *and when* $\mu \neq \underline{0}$, *the weighting vector maximising the power functions* (3.0.8) *is*

$$\omega^+ = \Sigma^{-1}\mu. \tag{3.1.1}$$

The proof of Theorem 1, as well as those of all the following results, are given in appendix $F$. Theorem 1 shows that the optimal weighting vector for the linear combination $z$ and $t$ tests is a standardization of the treatment effect. We will refer to $\beta_z^+$ and $\beta_t^+$ as representing the maximum value of the $z$ and $t$ test power functions attained under $\omega^+$. The next corollary compares $\beta_t^+$ to the power of Hotelling's $\mathrm{T}^2$ test.

**Corollary 1.** *Under* (3.0.1), $\beta_t^+ = \beta_t(\omega^+, \mu, \Sigma, n_y)$ *is larger or equal than the power of Hotelling's* $T^2$ *test* $\beta_{T^2}(\mu, \Sigma, n_y)$, *for any value of* $\mu$, $\Sigma$, $n_y > K$.

From Corollary 1 it is clear that, for an optimally selected weighting vector $\omega^+$, the linear combination $t$-test has higher power than Hotelling's $T^2$ for sample size $n_y > K$. Furthermore in section 4.3 we show that $\beta_{T^2}(\mu, \Sigma, n_y)$ takes considerably lower values than $\beta_t^+$ when $n_y$ approaches $K$ from above, i.e. $n_y \gtrapprox K$. Therefore, there is scope for preferring the linear combination testing procedures studied in this work especially when $n_y \gtrapprox K$ or $n_y \leq K$, that is when the $T^2$ test tends to lack power or is not applicable.

## 3.2 Selecting the weighting vector

Since the optimal weighting vector $\omega^+$ depends on the unknown model parameters $\mu$ and $\Sigma$, suitable estimates must be selected to actually implement such testing procedures. In this work, we propose estimating the weighting vector using prior information and data collected from a *pilot study* [25] conducted prior to the main study. Parsimonious modeling assumptions linking the moments of the pilot data $X$ with those of the main study data $Y$ are made. Assuming $X$ and $Y$ are sampled under the same conditions we let the $K$-dimensional pilot observations be conditionally independent Gaussian random variables

$$X_i \overset{\text{iid}}{\sim} \mathcal{N}_K(\theta, \Sigma), \; i = 1, 2, ..., n_x (n_x \geq 0) \tag{3.2.1}$$

with mean $\theta = (\theta_1, ..., \theta_K)'$ and covariance matrix $\Sigma$. While the covariance matrix $\Sigma$ is assumed to be the same in the two studies, we allow for heterogeneity between the mean responses. In particular, $\mu$ and $\theta$ are linked by a Gaussian transfer distribution

$$(\mu|\Sigma, \theta) \sim \mathcal{N}_K(\theta, \Sigma/d), \tag{3.2.2}$$

where $1/d$ represents the discrepancy between $\theta$ and $\mu$. Prior information $D_0 = [\theta_0, n_0, S_0, \nu_0]$ elicited from previous studies and experts clinical opinion is used to inform standard conjugate multivariate priors for the response mean and covariance matrix. For the mean response we use the prior

$$(\theta|\Sigma, D_0) \sim \mathcal{N}_K(\theta_0, \Sigma/n_0), \tag{3.2.3}$$

where $\theta_0$ represents a prior estimate of the value of $\theta$ and $n_0$ corresponds to the number of observations that this prior estimate is based on. When $\Sigma$ is unknown, we use the inverse-Wishart $K \times K$ prior

$$(\Sigma|D_0) \sim \mathcal{IW}_{K\times K}\left(\nu_0, S_0^{-1}\right), \tag{3.2.4}$$

where $\nu_0$ and $S_0$ respectively represent the degrees of freedom and the scale matrix of the inverse-Wishart prior.

Under this standard Bayesian model, the posterior distribution of $\mu$ given prior information and pilot data $D_1 = \{x, D_0\}$ is $K$-dimensional Gaussian with covariance matrix $\Sigma\left(\frac{1}{n_0+n_x} + \frac{1}{d}\right)$ and mean $m_1 = \left(m_{1,1}, ..., m_{1,K}\right)$ with entries

$$m_{1,k} = \frac{n_0}{n_0 + n_x}\theta_{0,k} + \frac{n_x}{n_0 + n_x}\overline{x}_k, \tag{3.2.5}$$

where $\overline{x}_k = \frac{1}{n_x}\sum_{i=1}^{n_x} x_{k,i}$ is the $k$-th component of the multivariate sample mean of the pilot data. The conditional posterior distribution of the covariance matrix $(\Sigma|D_1)$ is inverse-Wishart with scale matrix

$$S_1 = S_0 + (n_x - 1)S_x + \frac{n_0 n_x}{n_0 + n_x}(\overline{x} - \theta_0)(\overline{x} - \theta_0)', \tag{3.2.6}$$

where $S_x$ is the sample covariance matrix of the pilot data (see [13]).

In particular, the modeling assumptions (3.2.1), (3.2.2), (3.2.3) and (3.2.4) encompass the following special cases: (i) no prior information is available or $n_0, \nu_0 \to 0$ , (ii) no discrepancies are expected between the two studies or $d \to \infty$ implying $\mu = \theta$ (iii) no pilot data are used to select the weighting vector or $n_x = 0$.

### 3.2.1 Predictive power function

The key advantage to using Bayesian parameter estimation is that full posterior distributions can be used for the selection of the weighting vector as opposed to conditioning on point estimates. To this end we introduce the notion of predictive power [48], which in the present context is defined by averaging (3.0.8) with respect to the posterior distributions of the model parameters. The predictive power for the $z$-test is

$$B_{D_1}(w, n_y) = E_{\mu|D_1}(\beta_z(w, \mu, n_y)) = Pr\left(|Z_w| > z_{\alpha/2}|D_1\right). \tag{3.2.7}$$

The following theorem gives the form of the weighting vector maximising (3.2.7) under the above conditional posterior distribution for $\mu$.

**Theorem 2.** *Under* (3.0.1), (3.2.1), (3.2.2) *and* (3.2.3), *the weighting vector maximising* (3.2.7) *is*

$$w_z^\star = \Sigma^{-1} m_1. \tag{3.2.8}$$

In Appendix $B$ we show that (3.2.8) maximises the expected probability of rejecting $H_0$ for any value $\mu \neq 0$, that is the expected probability of rejecting $H_0$ given that it is false. Using the same steps followed in the proof of Theorem 2, in Appendix $G$ we also show that for sufficiently large $\nu_1 = \nu_0 + n_x$ the predictive power of the $t$-test

$$B_{t;D_1}(w, n_y) = E_{\mu,\Sigma|D_1}(\beta_t(w, \mu, \Sigma, n_y)) = Pr\left(|T_w| > t_{n_y-1,\alpha/2}|D_1\right), \tag{3.2.9}$$

is maximised by the weighting vector

$$w_t^\star = S_1^{-1} m_1. \tag{3.2.10}$$

Herein we will refer to $z^\star$ and $t^\star$ as representing the linear combination $z$ and $t$ test statistics with weighting vectors $w_z^\star$ and $w_t^\star$, respectively. It is important to note that $w_z^\star$ and $w_t^\star$ are chosen before the initiation of the main study based only on pilot sample data and/or prior estimates and therefore the $z^\star$ and $t^\star$ tests control the false positive rate at their nominal significance level.

Under (3.2.8) and (3.2.10), response components with large expected treatment effects and small variances receive larger weights. The weighting vectors $w_z^\star$ and $w_t^\star$ are also invariant to scale transformation of the response. Finally, it is the relative and not the absolute size of the treatment effects across the response components; precisely it is the *direction* of the global treatment effect, that determines the value of $w_z^\star$ and $w_t^\star$ as well as the optimal $\omega^+$. We express the latter result formally for 2- and 3-dimensional responses in corollary 2, Appendix $C$.

Using (3.2.8) and (3.2.10) it is easy to derive the optimal weighting vectors under different prior hyperparameters. First, when Jeffrey's priors ("noninformative" priors, see [13]) are used the optimal weighting vectors for the $z^\star$- and $t^\star$-test are respectively $w_z^\star = \Sigma^{-1}\overline{x}$ and $w_t^\star = S_x^{-1}\overline{x}$. Second, the values of the optimal weighting vectors (3.2.8) and (3.2.10) are independent of the assumed discrepancy between pilot and main study data $d$. Third, when no pilot data are available, that is $n_x = 0$, the optimal weighting vectors are respectively $w_z^\star = \Sigma^{-1}\theta_0$ and $w_t^\star = S_0^{-1}\theta_0$. In this case by taking $\theta_0 = m1_K$, when the variances of all response components are equal and all covariances are equal the optimal weighting vectors $w_z^\star$ and $w_t^\star$ coincide with the OLS solutions in [35] (see Corollary 3 Appendix C).

# 4 Power analysis

In this section we investigate the properties of the power functions

$$\beta_z^\star = \beta_z(w_z^\star, \mu, n_y), \tag{4.0.11}$$
$$\beta_t^\star = \beta_t(w_t^\star, \mu, \Sigma, n_y), \tag{4.0.12}$$

with respect to the model parameters $\mu, \Sigma, n_x/n_T$, $n_T = n_x + n_y$ using simulated and real fMRI data.

In Appendix D we show that, for $f = n_x/n_T > 0$, $\beta_z^\star$ and $\beta_t^\star$ are *consistent* [27], that is both power functions tend to one as $n_T \to \infty$. In Appendix D we also prove that for $f > 0$ the $z^\star$- and $t^\star$-tests are strictly unbiased with probability 1. Note that this is not generally the case for the linear combination $z$- and $t$-tests with fixed weighting vectors. For instance, the tests in [35] using the weighting vector $w^{OLS} = 1_k$ are biased because their power is not ensured to be larger than the significance level for $\mu \neq 0$. For instance, when $\mu = (-1, 1)'$ then $(w^{OLS})'\mu = 0$ and hence the power of the $z$- and $t$-tests using $w^{OLS}$ coincide with their significance level.

Prior to observing the pilot data $X = x$, the optimal weights $w_z^\star, w_t^\star$ and the power functions $\beta_z^\star$, $\beta_t^\star$ are random variables. To investigate their properties numerically we use a set of simulation studies varying the values of the model parameters $\theta$, $\mu$, $\Sigma$, $n_T$ and $f$ and the hyperparameters $\theta_0$, $n_0$, $S_0$, $\nu_0$ and $d$. Setting these studies prior to observing $x$ enables us to examine the impact of the suggested method for selecting the weighting vectors on the power of the linear combination $z$- and $t$-tests. For each set of parameter values we generate $r = 1, ..., R = 15000$ $K$-dimensional ($K = 4, 11$) synthetic datasets as follows:

1. Sample $\mu$ from $\mathcal{N}_K(\theta, \Sigma/d)$.
2. Sample the pilot data $x_r = (x_{ri})_{i=1}^{n_x}$ from $\mathcal{N}_K(\theta, \Sigma)$.
3. Compute $w^\star_{z,r}$ using (3.2.8) (or $w^\star_{t,r}$ using (3.2.10)).
4. Compute $\beta^\star_{z,r} = \beta_z\left(w^\star_{z,r}, \mu, n_y\right)$ (or $\beta^\star_{t,r} = \beta_t\left(w^\star_{t,r}, \mu, \Sigma, n_y\right)$).

In the following figures we depict simulation-based approximations of selected percentiles[2] of the distributions of $\beta^\star_z$ and $\beta^\star_t$. The compound-symmetric (CS) covariance matrix, with equal variances $\sigma^2$ and correlations $\rho$ across the response components, is used in the simulation studies so that variations in the strength of the treatment effect are entirely due to the structure of $\mu$. Note that by Theorem 1 the power functions of the $z$- and $t$-test attain their maxima if $w^\star_z$ and $w^\star_t$ are respectively equal to the weighting vector $\omega^+$ defined in (3.1.1). Power values lower than $\beta^+_z$ and $\beta^+_t$ are due to suboptimal selection of the weighting vector.

## 4.1 Sample Size

This section investigates how the sample sizes of pilot and main study, $n_x$ and $n_y$, affect the power of the $z^\star$- and $t^\star$-test under four scenarios: (i) $n_x$ only varies, (ii) $n_y$ only varies, (iii) the total sample size $n_T = n_x + n_y$ varies for fixed allocation ratios, $f = n_x/n_T$, of samples to pilot and main study (iv) the sample allocations $f$ vary for a fixed total sample size $n_T$. The first two scenarios characterise the behavior of the power functions when no constraints are imposed on the fMRI experimental design. The third scenario considers the typical question of sample size selection while the latter scenario addresses the issue of optimal sample size allocation when, for instance due to budget or capacity constraints, only a fixed number of subjects can be sampled. Note that the quantiles of $\beta^\star_t$ lies at lower levels than those of the power function of the $z^\star$-test, $\beta^\star_z$. This is due to the incorporation of extra uncertainty about the response's covariance matrix $\Sigma$.

The left panel in figure 2 shows that higher $n_x$ values result in higher values of the power functions $\beta^\star_z$ and $\beta^\star_t$. Also, since power is a non-linear mapping of the sample summaries, the distribution of the power function becomes progressively more skewed to the left as the pilot sample size increases. This behavior reflects an increase in precision of $\overline{x}, S_x$ and of $w^\star_z$ and $w^\star_t$, as estimators of $\omega^+$, as $n_x$ grows. The right panel in figure 2 shows that increases of $n_y$, similarly result in higher power and in a progressively more skewed power distribution. Note that, unlike for increases of $n_x$, the precision of the power function does not substantially increase as $n_y$ increases. This different behavior reflects the fact that the uncertainty of the power functions is introduced only by the pilot data through the weighting vectors.

Due to the consistency of the $z^\star$- and $t^\star$-test statistics, figure 3 (left panel) shows that increasing values of $n_T$ result in progressively more concentrated and higher power values. In particular, when using these simulated data, total sample sizes as low as $n_T = 15$ ensure that the median power is at least 0.85. Therefore, to the extent that these simulations resemble real fMRI data, the proposed test can attain high power for sample sizes typical of phase II clinical trials. The right panel of figure 3 illustrates the changes of $\beta^\star_z$ and $\beta^\star_t$ power functions with respect to $f$. If $n_x$ is small relatively to $\sigma$, the data $x$ may lead to a wrong selection of the weighting vector resulting in low power while for small $n_y$ power values also decrease. This simulation study suggests that under the present modeling parameters a range $f \in [0.3, 0.5]$ ensures attaining the maximum detection power for global treatment effects, whereas values $f < 0.3$ should only be considered when reliable prior information is available. The more precise the prior estimates are, the smaller $f$ values should be selected to gain advantage in terms of power. In supplementary material of this paper, we study the relation between the power functions and the allocation ratio for varying total sample size and prior information.

## 4.2 Heterogeneity between pilot and main studies

The simulation results displayed so far assume no discrepancies in mean between the pilot and main studies, that is $d \to \infty$ in (3.2.2). As shown in figure 4, when the pilot and main study samples are generated using low values of $d$, i.e. high discrepancies, and Bayesian estimates are calculated under Jeffrey's prior, the power distributions are concentrated at relatively low values. The power of the $z^\star$- and $t^\star$-tests and especially the upper percentiles of $\beta^\star_z$ and $\beta^\star_t$ (50-th,75-th, etc.) increase quickly with $d$ and all power functions are then unchanged once $d$ exceeds a small fraction of the total sample size, here approximately $n_T/10$. These results indicate that, even for moderate total sample sizes, $\beta^\star_z$ and $\beta^\star_t$ are relatively robust with respect to a small heterogeneity among the mean of the pilot and main study data. On the other hand, the power functions decrease in pathological situations, when the two sets of samples are recorded under substantially different conditions.

[2]Note that if the $p-th$ percentile of $\beta^\star_z$ is equal to $b$, then $Pr(\beta^\star_z \leq b) = p$.

### 4.3 Power comparison with other testing procedures

We next compare the $t^\star$-test with Hotelling's $T^2$ as well as the OLS $t$-test of [35] and the SS and PC $t$-tests of [26]. First note that by Corollary 1 the optimal power of the $t$-test, $\beta_t^+$, is larger or equal than $\beta_{T^2}$ for any sample size for which the $T^2$ test can be evaluated (i.e. $n_y > K$). Furthermore, $\beta_{T^2}$ is considerably lower than $\beta_t^+$ when $n_y$ approaches $K$ from above. For example, if $\mu = 0.5 \times e_{1,2,3}$[3], $\Sigma$ is CS with $\sigma = 0.5$ and $\rho = 0.5$, $n_y = 7, 8, 9, 10$, then $\beta_t^+(n_y) = 0.98, 0.99, 0.99, 0.99$ and $\beta_{T^2}(n_y) = 0.11, 0.25, 0.43, 0.61$ respectively. The power of the $t$-test can be higher than that of the $T^2$ test even when $n_y$ is well above $K$. For example, if $\mu = 0.5 \times 1_5$, $\Sigma$ unchanged and $n_y = 17, 18, 19, 20$, then $\beta_t^+(n_y) = 0.72, 0.74, 0.77, 0.79$ and $\beta_{T^2}(n_y) = 0.30, 0.33, 0.36, 0.38$ respectively. In Appendix E we report a study of the behaviour of the power function and of the connection between the optimal $z$- and $t$-tests and Hotelling's $T^2$ test through Mahalanobis distance. In the supplementary material, we describe Lauter's SS and PC $t$-tests.

In practice, the optimal weighting vector $\omega^+$ is unknown and its estimation relies on prior information and a pilot sample, which is likely to be small due to ethical and economical constraints. This suggests that prior estimates of the treatment effect can be very influential. In Table 1, we report the power functions $\beta_t^\star$, $\beta_{T^2}$, $\beta_t^{OLS}$ of the OLS $t$-test and $\beta_t^{SS}$, $\beta_t^{PC}$ of SS and PC $t$-tests, respectively, calculated for a simulated ROI fMRI study with $K = 11$, $n_x = 5$ and $n_y = 15$ generated using the algorithm described above. The power functions $\beta_{T^2}$, $\beta_t^{OLS}$, $\beta_t^{SS}$ and $\beta_t^{PC}$ are calculated using the total sample size $n_T = 20$ while the median $\beta_{t,50}^\star$ of the distribution of $\beta_t^\star$ at sample size $n_y = 15$. In other words, we suppose that the pilot study is undertaken with a cost of 5 observations to the available total sample size. We fix the covariance structure to being CS with $\sigma^2 = 0.05$ and high correlations $\rho = 0.65$. The hyperparameters for the prior distributions are taken as $\nu_0 = 1$, $S_0$ CS with $s_0^2 = 0.1$, $r_0 = 0.7$, $1/d = 0$ while $n_0 = 1$ and $\theta_0$ vary as shown in Table 1. We fix the Mahalanobis distance $\sqrt{\mu'\Sigma^{-1}\mu} = 0.90$ so that for all mean structures $\mu$ used in table 1), $\beta_t^+(n_y) = 0.90$, $\beta_t^+(n_T) = 0.97$ and $\beta_{T^2}(n_T) = 0.38$.

Table 1 shows that $\beta_t^\star$ can take substantially higher values than $\beta_{T^2}$ even for fairly poor priors and pilot estimates. For such priors, $t^\star$ is also considerably more efficient than the OLS, SS and the PC test, unless the treatment effect across the ROI is uniform (lines 1,2,3), in which case, as we prove in Corollary 3, $w^{OLS} = \omega^+$. Note that the behaviour of the latter power functions is rather similar. Differences in power of $t^\star$ with OLS, SS and PC tests are especially large when opposing effects are exhibited across ROI. The power of the latter tests is close to the significance level for opposing effect structures as in lines 13,15. Fairly precise prior information can substantially improve $\beta_t^\star$ while less informative or misleading priors decrease the value of $\beta_t^\star$. In line 3 of table 1, where the prior distribution for $\mu$ is "uninformative", i.e. $n_0 = 0$, the lowest presented value of $\beta_{t,50}$ is observed. Further, in lines 4,7,9,11,13,15, where the uniform prior estimate $\theta_0 \propto 1_{11}$ is used while the true treatment effect structures is in fact non-uniform, $\beta_{t,50}$ is substantially decreased. On the other hand, the more precise values of $\theta_0$ used in lines 2,5,6,8,10,12,14,16, result in up to 80% increase of the value of $\beta_{t,50}$ compared to uniform $\theta_0$.

#### 4.3.1 An example using a real fMRI study

A total of 11 subjects participated in a GSK study informing drug development using fMRI. At the planning stage, the following ROI were defined: 1. Anterior Cingulate (AC), 2. Atlas Amygdala (A), 3. Caudate (C), 4. Dorsolateral Prefrontal Cortex (DLPFC), 5. Globus Pallidus (GP), 6. Insula (I), 7. Orbitofrontal cortex (OFC), 8. Putamen (P), 9. Substantia Nigra (SA), 10. Thalamus (T), 11. Ventral Striatum (VS). We use the values of a linear contrast between treatment and placebo for each ROI and each subject (see table 2) for testing the global null hypothesis of no treatment effect. Note that substantially different effect sizes are observed across ROI while the sample correlations are generally high. In particular, the correlations between DLPFC and C, in both pilot and main study data, are especially high. Note also the relatively large differences between the observed sample statistics of the pilot and main study data.

We implement $t^\star$-test using observations from the first three subjects to select the weighting vector, i.e. $n_x = 3$, $n_y = 8$. This corresponds to $f \lessapprox 0.3$ which, as we discussed in section 4.1, is preferable in terms of power for fairly vague prior distributions. For comparison, we also implement OLS, SS and PC tests using the whole sample $n_T = 11$. Hotelling's $T^2$ test cannot be implemented since the number of ROI $K = 11$ is equal to the total sample size. As shown in table 3, even for vague prior estimates, $t^\star$ can attain lower p-values (see lines 2-7,9-14) and reject the null hypothesis at significance level $\alpha = 0.05$ (see lines 6,7,13,14), unlike the other tests. In lines 1-7 of table 3 we suppose that the investigators *a-priori* believe that the variance-covariance matrix is CS while in lines 8-14 the belief that the ROI C and DLPFC are highly correlated is incorporated into the prior of $\Sigma$ resulting to a substantial decrease of $p_{t^\star}$-values. This is in stark contrast with the results reported at lines 15 and 16, where due to incorrect prior information, lower prior correlations are taken between the same ROI and $p_{t^\star}$ increases, failing to reject the null hypothesis. Decrease of the value of $p_{t^\star}$ is also succeeded in lines 4-7,

[3] $e_{i_1,\ldots,i_l}$ is a $K$-dimensional vector with non-zero entries only at indices $i_1, \ldots, i_l$, e.g. $e_{1,2,3} = (1, 1, 1, 0, 0, \ldots, 0)'$

11-14 and 16 where the prior estimates of the treatment effects in the ROI T and VS are set twice larger than the other effects. On the other hand, by taking $n_0 = 0$, $p_{t^\star}$ is substantially larger than $p_{OLS}$, $p_{SS}$, $p_{PC}$ (see lines 1,8). These results emphasize the strong effect of prior information when the total sample size is small and the improvement of the performance of the linear combination $t$-test if precise prior information is available.

## 5 Discussion and Further Extensions

We construct a procedure for testing the global null hypothesis of no treatment effect against its two-sided alternative based on linear combinations of multivariate responses. These linear combinations, derived using optimally selected weighting vectors, are used for constructing the classical $z$ and $t$ statistics. In corollary 1, we show that these linear combination tests can achieve higher power than the classical Hotelling's $T^2$ test. The difference between the power of the two tests is particularly large when the sample size of the study is close to the dimension of the multivariate response. We emphasize the relevance of the developed tests under the typical sample size restrictions for the motivating fMRI ROI studies.

Our proposals generalises the linear combination tests proposed by O'Brien [35] and Lauter et al. [26] by allowing for efficient detection of treatment effects of general structures. The proposed $z^\star$- and $t^\star$-tests are particularly useful when at the planning stage of clinical trials the investigators believe that the structures required in [35] and [26] for efficient detection of the global treatment effects are inappropriate. Typically, at this stage, there is a lack of precision of the required estimates and therefore combining the available information with pilot data for selecting the weighting vector appears to be a potentially preferable option.

We also extend the use of the predictive power function to the selection of the weighting vector. Predictive power has been used for sample size calculation [47], [36], [20], interim analysis [48], treatment selection [24] and to select the component-wise significance levels in multiple testing [53]. It is a useful tool in hybrid frequentist-Bayesian approaches where the predictive power function is used for designing the trial and frequentist analysis controlling the false positive rates is planned for the end of the trial [46].

Ours is also a two-stage procedure, where the first stage pilot data is used for selecting the weighting vector, while the second stage main study data are used to undertake the hypothesis test. Pilot data are commonly collected for designing larger main studies [25], [44]. If these pilot data are collected at the same conditions with the main study, they can be used for selecting the weighting vector. However, as we have seen in the simulation study in section 4.3, the $t^\star$-test can succeed larger power than Hotelling's $T^2$, OLS, SS and PC tests even in cases where a pilot study has not been conducted and a part of the sample is used for selecting the weighting vector.

For small pilot sample sizes, the efficiency of the suggested testing procedure largely relies on prior estimation of the unknown parameters. However, for sufficiently precise prior estimates, the proposed $z^\star$- and $t^\star$-tests can attain higher values than Hotelling's $T^2$, OLS, SS and PC tests. Furthermore, if the sample allocation to the two stages is positive, the proposed $z^\star$- and $t^\star$-tests are consistent and, unlike to OLS test, unbiased with probability 1. The $z^\star$- and $t^\star$ are also scale invariant with respect to the response and the weighting vector. The distribution of the test statistic under the alternative hypothesis is known and therefore it can be studied analytically.

In many occasions, investigators may wish to undertake a testing procedure where the alternative hypothesis is more constrained. A potential extension of the suggested procedure is to modify accordingly the alternative hypothesis and the selection of the weighting vector by imposing the appropriate constraints. Another development path is to extend the use of the pilot data by including them in the final dataset for testing. In this case, caution should be taken so that the false positive rate is controlled. Adaptive methods might be appropriate in this case. If all available data are fully used for testing without inflating the type I error rate, one may wish to extend the proposed procedures in more than two stages, where the weighting vectors are modified using the accumulated information at each interim analysis. This is an area of ongoing research.

We would like to thank Paul M. Matthews of GlaxoSmithKline for motivating this work and providing the data.

# Appendices

## A The distribution of the test statistics

By Theorem 3.3.2 p.71 and Theorem 3.3.1 p.68 in Anderson [2], the test statistic $Z_w \sim \mathcal{N}(\delta_w, 1)$, where

$$\delta_w = \frac{w'\mu}{\sigma_L/\sqrt{n_y}}. \tag{A.0.1}$$

Under the null hypothesis in (3.0.3), for any fixed weighting vector $w \neq 0$, $Z_w \sim \mathcal{N}(0,1)$ and hence the $z$-test in (3.0.6) controls the false positive rate.
Using Theorem 3.3.1 p.68 in [2], we can see that the sample mean of the linear combination $L$ is $\overline{L} \sim \mathcal{N}\left(w'\mu, w'\Sigma w/n_y\right)$. Further, by Corollary 7.2.3 p.249 in [2] the sample covariance matrix $S_y$ of $Y$ follows the Wishart distribution with parameters $\Sigma$ and $n_y - 1$. Thus, using Theorem 3.4.2 p.67 in Mardia et al. [31] it can be shown that $\frac{(n_y-1)s_l^2}{w'\Sigma w} \sim \chi^2_{n_y-1}$ where $s_l^2$ is the sample variance of $L$. Hence, the $t$-statistic $T_w$ can be written as

$$T_w = \frac{Z + \delta_w}{\sqrt{X/(n_y-1)}},$$

where $Z \sim \mathcal{N}(0,1)$ and $X \sim \chi^2_{n_y-1}$ and thus it is noncentrally t-distributed with noncentrally parameter $\delta_w$ and $n_y - 1$ degrees of freedom. Under the null hypothesis in (3.0.3), for any fixed $w \neq 0$, $T_w \sim t_{n_y-1}$ and hence the $t$-test in (3.0.6) controls the false positive rate.

# B Conditional and unconditional predictive power

Using probability law, we can see that

$$\begin{aligned} Pr\left(|Z_w| > z_{\alpha/2}|D_1\right) &= Pr\left(|Z_w| > z_{\alpha/2}, H_0 \text{ true}|D_1\right) + Pr\left(|Z_w| > z_{\alpha/2}, H_0 \text{ false}|D_1\right) \\ &= Pr\left(|Z_w| > z_{\alpha/2}|H_0 \text{ true}, D_1\right) Pr(H_0 \text{ true}|D_1) + Pr\left(|Z_w| > z_{\alpha/2}|H_0 \text{ false}, D_1\right) Pr(H_0 \text{ false}|D_1). \end{aligned}$$

Note that $Pr(H_0 \text{ true}|D_1)$ and $Pr(H_0 \text{ false}|D_1)$ are independent of the weighting vector $w$ and $Pr\left(|Z_w| > z_{\alpha/2}|H_0 \text{ true}\right) = \alpha$ for any fixed $w \neq 0$. Therefore, maximisation of either $Pr\left(|Z_w| > z_{\alpha/2}|D_1\right)$ or $Pr\left(|Z_w| > z_{\alpha/2}|H_0 \text{ false}, D_1\right)$ with respect to $w$ is equivalent.

# C Direction of $\omega^+$, $w_z^\star$ and $w_t^\star$

The next corollary describes the factors determining $\omega^+$.

**Corollary 2.** *The weighting vector $\omega^+$, for a $K$-dimensional response, $K = 2, 3$ is completely determined by (i) the ratios*

$$m_{ij} = \frac{\mu_i}{\mu_j} \text{ and } s_{ij} = \frac{\sigma_{ii}}{\sigma_{jj}}, \tag{C.0.2}$$

*and (ii) the correlations*

$$\rho_{ij} = \sigma_{ij}/\sqrt{\sigma_{ii}\sigma_{jj}}, \tag{C.0.3}$$

*for all $i, j \in \{1, ..., K\}$, $i \neq j$.*

The same results can be proved for $w_z^\star$ (and $w_t^\star$) if $\mu$ (and $\Sigma$) is replace by $m_1$ in (3.2.5) (and $S_1$ in (3.2.6)).

**Corollary 3.** *The optimal weighting vector $\omega^+$ is equal to $w^{GLS}$ if $\mu = m1_k$ and to $w^{OLS}$ if $\mu = m1_k$ and $\Sigma$ is compound symmetric.*

The same results can be proved for $w_z^\star$ (and $w_t^\star$) if $\mu$ (and $\Sigma$) is replaced by $m_1$ in (3.2.5) (and $S_1$ in (3.2.6)). This corollary support the suggestions of Sankoh et al. [42] and Neuhauser [34] considering the power of the OLS and GLS tests in [35].

# D Properties of the testing procedure

## D.1 Unbiasedness

A test is strictly unbiased if the power of the test is greater than the significance level for any treatment effect $\mu \neq 0$. For $\mu \neq 0$, the power functions $\beta_z^\star$ and $\beta_t^\star$ exceeds the significance level $\alpha$ if and only if $\delta_{w_z^\star}$ or $\delta_{w_t^\star}$, respectively, is non-zero. By allowing the weighting vectors $w_z^\star$ and $w_t^\star$ and thus $\delta_{w_z^\star}$ and $\delta_{w_t^\star}$ to depend on the pilot observations, they become continuous random variables and therefore the probability of the events described above is 1. In summary, for $f = n_x/n_T > 0$ and $\mu \neq 0$, both $\beta_z^\star$ and $\beta_t^\star$ are greater than $\alpha$ with probability 1.

### D.2 Consistency

From the weak law of large numbers, $\overline{x}$ and $S_x$ converge in probability to $\mu$ and $\Sigma$, respectively, as $n_x \to \infty$. Therefore, $w_z^\star$ and $w_t^\star$ converge in probability to $\omega^+ = \Sigma^{-1}\mu$ which implies that both $\sqrt{n_y}\delta_{w_z^\star}$ and $\sqrt{n_y}\delta_{w_t^\star}$ converge in probability to the Mahalanobis distance $\sqrt{\mu'\Sigma^{-1}\mu}$ (see appendix E), as $n_x \to \infty$. Since, $\Sigma$ is positive definite, $\sqrt{\mu'\Sigma^{-1}\mu}$ is positive for any $\mu \neq 0$. Therefore, for any $\mu \neq 0$ and $f > 0$, $\delta_{w_z^\star}$ and $\delta_{w_t^\star}$ both converge in probability to infinity, as $n_T \to \infty$ which implies that, under these conditions, the power functions $\beta_z^\star$ and $\beta_t^\star$ both converge in probability to 1.

## E Connection between linear combination $z$- and $t$-test and Hotelling's $T^2$ test

The linear combination $z$- and $t$-tests are connected with Hotelling's $T^2$ test through Mahalanobis distance (MD) $D_{\mu,\Sigma} = \sqrt{\mu'\Sigma^{-1}\mu}$. The noncentrality parameter of the noncentral $F$ distribution of Hotelling's $T^2$ statistic is $n_y \left(D_{\mu,\Sigma}\right)^2$ while from Theorem 1 we have that the mean and the noncentrality parameter of the $z$- and $t$-statistics, respectively, attained under $\omega^+$, are equal to $n_y D_{\mu,\Sigma}$. We can explore the power of these tests as functions of $n_y$ and $D_{\mu,\Sigma}$. MD is a non-negative measure of the discrepancy between the density of the distribution $\mathcal{N}_K(\mu,\Sigma)$ and the density of the null distribution $\mathcal{N}_K(\underline{0},\Sigma)$. All other parameters being equal, larger discrepancies of the former distribution to the latter result in higher power values. For instance, if $\Sigma$ is compound symmetric with high correlations, the distribution $\mathcal{N}_K(1_K,\Sigma)$ is more distant from the null distribution than the distribution $\mathcal{N}_K(e_1,\Sigma)$ and therefore $D_{1_K,\Sigma} < D_{e_1,\Sigma}$ which in turn implies that $\beta_{T^2}(1_K,\Sigma,n_y) < \beta_{T^2}(e_1,\Sigma,n_y)$ and $\beta_t^+(1_K,\Sigma,n_y) < \beta_t^+(e_1,\Sigma,n_y)$.

## F Proofs

*Proof of Theorem 1.* The power functions $\beta_z(w,\mu,n_y)$ and $\beta_t(w,\mu,\Sigma,n_y)$ are maximised with respect to $w$ if and only if $|\delta_w|$ or equivalently $\delta_w^2$ is maximised. Using the Generalised Cauchy-Schwartz inequality (see p. 178 [2]) we have that $\delta_w^2 \leq \mu'\Sigma\mu$. Since $\delta_{\omega^+}^2 = \frac{(\mu'\Sigma^{-1}\mu)^2}{\mu'\Sigma^{-1}\Sigma\Sigma^{-1}\mu} = \mu'\Sigma\mu$, $\omega^+$ attains the maximum of $|\delta_w|$. □

*Proof of Corollary 1.* Let $\mu$, $\Sigma$, $n_y > K$ be arbitrary values of the mean, the covariance matrix and the sample size. Under (3.0.1), the $t$-statistic in (3.0.5), $T_{\omega^+}$ for $w = \omega^+$ follows the non-central $t$ distribution with non-centrality parameter $\delta_{\omega^+} = \sqrt{n_y\mu'\Sigma^{-1}\mu}$ and $n_y - 1$ degrees of freedom. Hence, the square of the latter $t$-statistic, $T_{\omega^+}^2$ follows the non-central $F$ distribution with non-centrality parameter $\delta_{\omega^+}^2$ and $(1, n_y - 1)$ degrees of freedom. Therefore, the power of the $t$-test (3.0.6) $\beta_t(\omega^+,\mu,\Sigma,n_y) = Pr_{\mu,\Sigma}\left(T_{\omega^+}^2 > F_{1,n_y-1,\alpha}\right)$ where $F_{1,n_T-1,\alpha}$ is the $100(1-\alpha)$ percentile of the (central) F distribution with $1, n_T - 1$ degrees of freedom. Furthermore, from properties of the noncentral $F$-distribution we have that $Pr_{\mu,\Sigma}\left(T_{\omega^+}^2 > F_{1,n_T-1,\alpha}\right) \geq Pr_{\mu,\Sigma}\left(T_{\omega^+}^2 > F_{K,n_y-K,\alpha}\right) = \beta_{T^2}(\mu,\Sigma,n_y)$, for $K > 1$, where $\beta_{T^2}(\mu,\Sigma,n_y)$ is the power of the Hotelling's $T^2$ test. Therefore, for any value of $\mu, \Sigma, n_T > K$, $\beta_t(\omega^+,\mu,\Sigma,n_T) \geq \beta_{T^2}(\mu,\Sigma,n_T)$. □

*Proof of Theorem 2.* The $z$-statistic can be written as $Z_w = \delta_w + e$, where $e \sim \mathcal{N}(0,1)$ and $\delta_w$ is defined in $(A.0.1)$. Under (3.2.1), (3.2.2), (3.2.3), (3.2.4), $(\delta_w|D_1) \sim \mathcal{N}\left(\frac{w'm_1}{\sqrt{w'\Sigma w/n_y}}, n_y\left(\frac{1}{n_0+n_x} + \frac{1}{d}\right)\right)$. Thus, $(Z_w|D_1) \sim \mathcal{N}\left(\frac{w'm_1}{\sqrt{w'\Sigma w/n_y}}, n_y(\frac{1}{n_0+n_x} + \frac{1}{d}) + 1\right)$. The result is then proved using the same steps as in Theorem 1 where $\mu$ is replaced by $m_1$ $(m_1 \neq \underline{0})$. □

*Proof of Corollary 2.* The weighting vector $\omega^+ = \Sigma^{-1}\mu$, for K=2 and K=3 is respectively

$$\begin{pmatrix} 1 - \frac{m_{21}\rho_{21}}{\sqrt{s_{21}}} \\ \frac{m_{21}}{s_{21}} - \frac{\rho_{21}}{\sqrt{s_{21}}} \end{pmatrix}, \quad \begin{pmatrix} (1-\rho_{23}^2) + \frac{m_{21}(\rho_{13}\rho_{23}-\rho_{12})}{\sqrt{s_{21}}} + \frac{m_{31}(\rho_{12}\rho_{23}-\rho_{13})}{\sqrt{s_{31}}} \\ \frac{(\rho_{13}\rho_{23}-\rho_{12})}{\sqrt{s_{21}}} + \frac{m_{21}(1-\rho_{13}^2)}{s_{21}} + \frac{m_{31}(\rho_{12}\rho_{13}-\rho_{23})}{\sqrt{s_{21}}s_{31}} \\ \frac{(\rho_{12}\rho_{23}-\rho_{13})}{\sqrt{s_{31}}} + \frac{m_{21}(\rho_{12}\rho_{13}-\rho_{23})}{\sqrt{s_{21}}s_{31}} + \frac{m_{31}(1-\rho_{12}^2)}{s_{31}} \end{pmatrix}. \tag{F.0.1}$$

Since the expressions in $(F.0.1)$ depend only on the quantities described in (i) and (ii) the result of Corollary 2 follows. □

*Proof of Corollary 3.* The result for $w^{GLS}$ and $w^{OLS}$ are respectively proved by setting $\mu = m1_K$ and $\mu = m1_K$, $\Sigma$ CS in (3.1.1). □

# G Maximisation of the predictive power of the $t$-test

Using Bayes Theorem, under (3.0.1), (3.2.1), (3.2.2), (3.2.3), (3.2.4), we have that: (i) $(\Sigma|D_1) \sim \mathcal{IW}_{K\times K}(\nu_1, S_1^{-1})$, where $\nu_1 = \nu_0 + n_x$, $S_1$ as in (3.2.6) and $\overline{x}$ and $S_x$ the sample mean and covarince matrix of X, (ii) $\left(\overline{Y}|D_1\right) \sim t_K\left(m_1, NS_1/n_y, \nu_1 - K + 1\right)$, where $m_1$ is given in (3.2.5) and $N = \frac{2+n_0+n_x+d}{(n_0+n_x)d(\nu_1-K+1)}$. Using the results in [50] we derive the predictive distribution $\left(\overline{L}|D_1\right) \sim t\left(w'm_1, Nw'S_1w/n_y, \nu_1 - K + 1\right)$. For large $\nu_1$,

$$\left(\overline{L}|D_1\right) \simeq \mathcal{N}\left(w'm_1, Nw'S_1w/n_y\right). \tag{G.0.2}$$

By Bayes Theorem, $(Y_i|D_1) \sim t_K\left(m_1, NS_1, \nu_1 - K + 1\right), i = 1, 2, ..., n_y$. For large values of $\nu_1$, $(Y_i|D_1) \simeq \mathcal{N}\left(m_1, NS_1\right), i = 1, 2, ..., n_y$. By Corollary 7.2.3 p.249 in [2], it follows that $(S_y|D_1) \simeq \mathcal{W}\left(n_y - 1, NS_1/n_y - 1\right)$ and hence by Theorem 3.4.2 p.67 in Mardia et al. [31]

$$\left(\frac{(n_y-1)s_L^2}{Nw'S_1w}|D_1\right) \simeq \chi^2_{n_y-1} \ \ (Nw'S_1w > 0). \tag{G.0.3}$$

From $(G.0.2)$ and $(G.0.3)$, it follows that $(T_w|D_1)$ can be written as $T_w = \frac{Z+d_w}{\sqrt{X/(n_y-1)}}$ where $(Z|D_1) \simeq \mathcal{N}(0,1)$ and $(X|D_1) \simeq \chi^2_{n_y-1}$ and hence it is noncentrally t-distributed with noncentrality parameter $d_w = \frac{w'm_1}{\sqrt{Nw'S_1w/n_y}}$ and $n_y - 1$ degrees of freedom. Following the same steps of the proof of Theorem 1, by replacing $\delta_w$ with $d_w$ it is easily proved that for large $\nu_1$, the predictive power (3.2.9) is maximised with respect to $w$ if $w = w_t^\star$.

# H Figures and tables

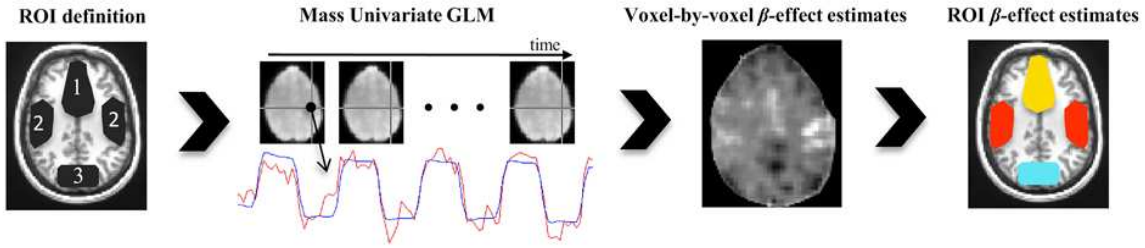
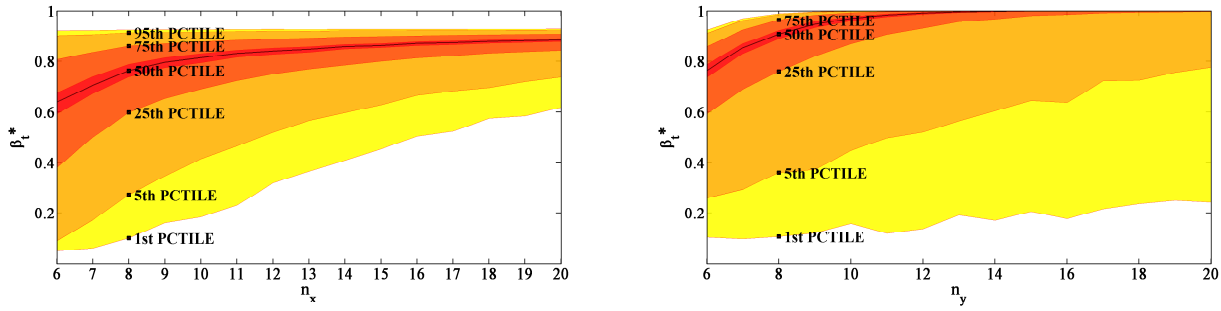


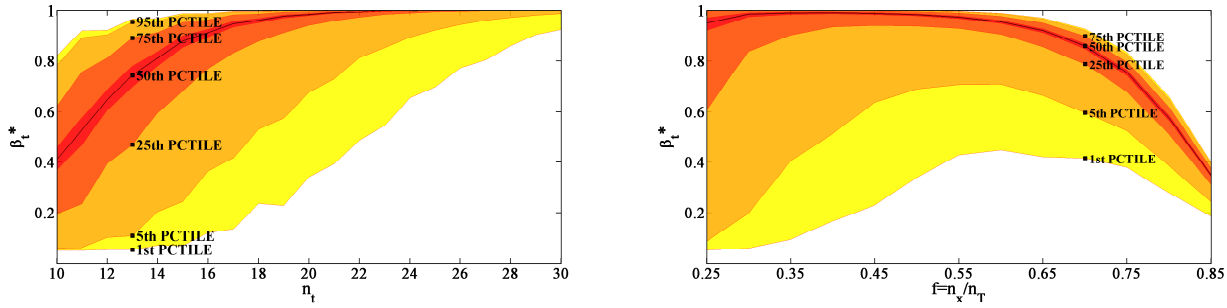

Figure 1: Typical steps of fMRI data analysis producing the multivariate outcome used in our methods. The preprocessed series of fMRI images are modeled at voxel-by-voxel resolution using mass univariate General Linear Models (GLMs). Suitable estimates of parameter values ($\beta$) expressing the treatment effect in each voxel are first extracted from the GLM and then averaged across the predefined ROI to produce the multivariate outcome we use to detect global treatment effects.




(a) $\beta_t^\star$ versus $n_x$ for $n_y = 6$

(b) $\beta_t^\star$ versus $n_y$ for $n_x = 8$

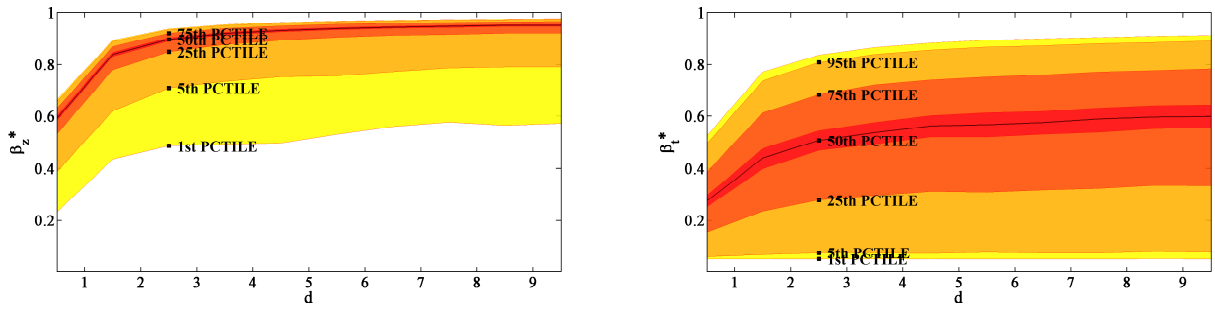
Figure 2: Simulation-based approximation of selected percentiles of the distributions of the power function $\beta_t^\star$ plotted against the pilot (left panel) and main study (right panel) sample sizes. The parameter values used here are: $\theta = (0.1, 0.1, 0.3, 0.3)'$, $1/d = 0$, $\sigma^2 = 0.05$, $\rho = 0.6$. Jeffrey's priors are used to derive Bayesian posterior estimates. Higher sample sizes increase the median power and the skewness of the power distribution. Higher pilot sample sizes result in a more concentrated power distribution due to their contribution both in estimating the weights $w$.




(a) $\beta_t^\star$ versus $n_T$ for $f = 0.5$

(b) $\beta_t^\star$ versus $f$ for $n_T = 20$

Figure 3: Simulation-based approximation of selected percentiles of the distributions of the power function $\beta_t^\star$ plotted against the total sample size of the two studies (left panel) and the sample allocation (right panel). The parameter values used here are: $\theta = (0.1, 0.1, 0.3, 0.3)'$, $\Sigma$ CS with $\sigma^2 = 0.05$ ,$\rho = 0.6$. Jeffrey's priors are used to derive Bayesian posterior estimates. As $n_T$ increases the distributions of the power function concentrates at higher levels. Higher power levels are achieved for $f \in [0.3, 0.5]$.




(a) $\beta_z^\star(X)$ (b) $\beta_t^\star(X)$

Figure 4: Simulation-based approximation of selected percentiles of the distributions of the power functions $\beta_z^\star$ (left panel) and $\beta_t^\star$ (right panel) plotted against $d$. The parameter values used here are $\mu = (0.1, 0.2, 0.3, 0.4)'$ , $\Sigma$ CS with $\sigma^2 = 0.1$ ,$\rho = 0.6$, $n_T = 14$, $f = n_x/n_y = 0.45$. As $d$ increases, the discrepancies between the two studies are reduced and the power function attains higher levels. The lower percentiles increase in lower rates than the upper percentiles.

Table 1: Power of OLS, SS, PC and $t^\star$-test (50-th percentile) for various hyperparameters, $\theta_0$ and $n_0$, and mean responses, $\mu = \frac{0.90}{\sqrt{\widetilde{\mu}'\Sigma^{-1}\widetilde{\mu}}} \times \widetilde{\mu}$. For such small sample sizes, the prior estimates of the treatment effect are highly influential on $\beta_t^\star$. Precise prior information results in substantially greater $\beta_t^\star$ compared to the other power functions presented here.

| | $\widetilde{\mu}$ | $\beta_t^{OLS}(n_T)$ | $\beta_t^{SS}(n_T)$ | $\beta_t^{PC}(n_T)$ | $\theta_0$ | $n_0$ | $\beta^\star_{t,0.50}(n_y)$ |
|---|---|---|---|---|---|---|---|
| 1 | $(1, 1, ..., 1)'$ | 0.97 | 0.96 | 0.96 | $0.25 \times 1_{11}$ | 1 | 0.60 |
| 2 | | | | | | 5 | 0.74 |
| 3 | | | | | | 0 | 0.33 |
| 4 | $(5, 1, 1, ..., 1)'$ | 0.16 | 0.14 | 0.14 | $0.25 \times 1_{11}$ | 1 | 0.37 |
| 5 | | | | | $0.25 \times e_1$ | 1 | 0.58 |
| 6 | | | | | | 5 | 0.68 |
| 7 | $(5, 5, 5, 5, 5, 1, 1, ..., 1)'$ | 0.20 | 0.20 | 0.20 | $0.25 \times 1_{11}$ | 1 | 0.39 |
| 8 | | | | | $0.1 \times e_{1,2,3}$ | 1 | 0.52 |
| 9 | $(5, 4, 3, 2, 1, 1, ..., 1)'$ | 0.19 | 0.18 | 0.18 | $0.25 \times 1_{11}$ | 1 | 0.39 |
| 10 | | | | | $0.1 \times e_{1,2,3}$ | 1 | 0.59 |
| 11 | $(6, 6, 4, 4, 2, 2, 1, 1, ..., 1)'$ | 0.28 | 0.26 | 0.27 | $0.25 \times 1_{11}$ | 1 | 0.42 |
| 12 | | | | | $0.1 \times e_{1,2} + 0.05 \times e_{3,4,5,6}$ | 1 | 0.60 |
| 13 | $(6, 6, 4, 4, 2, 2, -1, -1, ..., -1)'$ | 0.13 | 0.13 | 0.13 | $0.25 \times 1_{11}$ | 1 | 0.37 |
| 14 | | | | | $0.1 \times e_{1,2} + 0.05 \times e_{3,4,5,6}$ | 1 | 0.58 |
| 15 | $(-6, 6, 4, 4, 2, 2, 1, 1, ..., 1)'$ | 0.07 | 0.07 | 0.07 | $0.25 \times 1_{11}$ | 1 | 0.33 |
| 16 | | | | | $-0.1 \times e_1 + 0.05 \times e_{2,3,4,5,6}$ | 1 | 0.59 |

Table 2: Sample means (lines 1,2), sample variances (lines 3,4) and sample correlations (upper and lower diagonal of matrix in lines 5-15) of the pilot and main study data, $x$ and $y$, respectively. Opposing effects and generally high correlations are observed across ROI.

| 0 | ROI | AC | A | C | DLPFC | GP | I | OFC | P | SA | T | VS |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | $\overline{x}_k$ | 0.21 | 0.23 | 0.05 | -0.04 | -0.02 | 0.22 | -0.19 | 0.20 | 0.06 | 0.15 | 0.16 |
| 2 | $\overline{y}_k$ | -0.14 | -0.05 | -0.15 | -0.13 | -0.18 | -0.13 | -0.12 | -0.15 | -0.12 | -0.25 | -0.22 |
| 3 | $s_{x,k}$ | 0.54 | 0.36 | 0.18 | 0.36 | 0.49 | 0.30 | 0.62 | 0.59 | 0.09 | 0.47 | 0.44 |
| 4 | $s_{y,k}$ | 0.22 | 0.32 | 0.15 | 0.19 | 0.32 | 0.24 | 0.27 | 0.35 | 0.39 | 0.23 | 0.26 |
| 5 | AC | 1.00 | 0.97 | 0.99 | 0.99 | 0.90 | 0.97 | 0.94 | 0.99 | 0.70 | 0.99 | 0.79 |
| 6 | A | 0.40 | 1.00 | 0.99 | 0.95 | 0.96 | 0.99 | 0.99 | 0.98 | 0.54 | 0.98 | 0.90 |
| 7 | C | 0.78 | 0.20 | 1.00 | 0.98 | 0.93 | 0.99 | 0.96 | 0.99 | 0.63 | 0.99 | 0.84 |
| 8 | DLPFC | 0.78 | 0.34 | 0.96 | 1.00 | 0.85 | 0.94 | 0.89 | 0.99 | 0.77 | 0.99 | 0.73 |
| 9 | GP | 0.73 | 0.75 | 0.66 | 0.68 | 1.00 | 0.97 | 0.99 | 0.90 | 0.32 | 0.91 | 0.98 |
| 10 | I | 0.85 | 0.64 | 0.71 | 0.71 | 0.93 | 1.00 | 0.99 | 0.98 | 0.53 | 0.98 | 0.90 |
| 11 | OFC | 0.48 | 0.56 | 0.37 | 0.52 | 0.32 | 0.30 | 1.00 | 0.94 | 0.41 | 0.95 | 0.95 |
| 12 | P | 0.71 | 0.57 | 0.72 | 0.73 | 0.94 | 0.82 | 0.24 | 1.00 | 0.68 | 0.99 | 0.81 |
| 13 | SA | 0.32 | 0.79 | 0.19 | 0.36 | 0.60 | 0.49 | 0.32 | 0.51 | 1.00 | 0.67 | 0.13 |
| 14 | T | 0.86 | 0.33 | 0.81 | 0.84 | 0.70 | 0.70 | 0.58 | 0.77 | 0.19 | 1.00 | 0.81 |
| 15 | VS | 0.63 | 0.63 | 0.62 | 0.69 | 0.91 | 0.78 | 0.30 | 0.94 | 0.62 | 0.73 | 1.00 |

Table 3: P-values of $t^\star$-test for various hyperparameters $\theta_0$,$n_0$ and $S_0$ compared with the p-values of OLS, SS and PC tests computed using the data in table (2). Even for fairly poor prior estimates, $t^\star$- can succeed p-values lower than the $\alpha$-level.

| | $p_{OLS}$ | $p_{SS}$ | $p_{PC}$ | $S_0$ | $\theta_0$ | $n_0$ | $p_{t^\star}$ |
|---|---|---|---|---|---|---|---|
| 1 | 0.33 | 0.31 | 0.34 | CS:$s_0^2 = 0.05$, $r^0 = 0.6$ | $0.1 \times 1_K$ | 0 | 0.70 |
| 2 | | | | | | 1 | 0.29 |
| 3 | | | | | | 3 | 0.22 |
| 4 | | | | | $(0.1, ..., 0.1, 0.2)$ | 1 | 0.15 |
| 5 | | | | | | 3 | 0.10 |
| 6 | | | | | $(0.1, ..., 0.1, 0.2, 0.2)$ | 1 | 0.02* |
| 7 | | | | | | 3 | 0.01* |
| 8 | | | | same but $r^0_{3,4} = 0.9$ | $0.1 \times 1_K$ | 0 | 0.53 |
| 9 | | | | | | 1 | 0.20 |
| 10 | | | | | | 3 | 0.17 |
| 11 | | | | | $(0.1, ..., 0.1, 0.2)$ | 1 | 0.14 |
| 12 | | | | | | 3 | 0.11 |
| 13 | | | | | $(0.1, ..., 0.1, 0.2, 0.2)$ | 1 | 0.02* |
| 14 | | | | | | 3 | 0.01* |
| 15 | | | | same but $r^0_{3,4} = 0.2$ | $0.1 \times 1_K$ | 1 | 0.32 |
| 16 | | | | | $(0.1, ..., 0.1, 0.2)$ | 1 | 0.14 |

# I Supplementary Material

## I.1 Sample Allocation

In the following we explore how the power distribution of $\beta_t^\star$ is affected by different allocations of the total sample $n_T$ to the pilot and main study. We explore this effect for increasing values of $n_T$. In figure 5, where Jeffrey's priors are used, allocation ratios $f = n_x/n_T \in (0.3, 0.5)$ result in more concentrated power distributions that also attain higher levels. As $n_T$ increases, $\beta_t^\star$ distribution becomes more tight and achieve higher levels, but values of $f$ in the range of $(0.3, 0.5)$ remain superior. In figure 6, despite the fact that the conjugate priors used are fairly vague, the $\beta_t^\star$ distributions are generally changed. In particular, the upper percentiles (50-th, 75-th percentile etc.) of $\beta_t^\star$ in figure 6 are substantially higher than those in figure 5 since the percentage of wrong

selections of $w_t^\star$ is decreased. They also attain their higher values for small values of $f$, $f < 0.15$. On the other hand, the lower percentiles of $\beta_t^\star$ remain fairly unaffected compared to figure 5 since wrong selections of $w_t^\star$ are still observed due to inaccurate pilot estimates of $\mu$ and $\Sigma$. Similarly with Jeffrey's prior, although the power distribution concentrate in higher levels for large $n_T$, the range of most suitable $f$ remains the same.

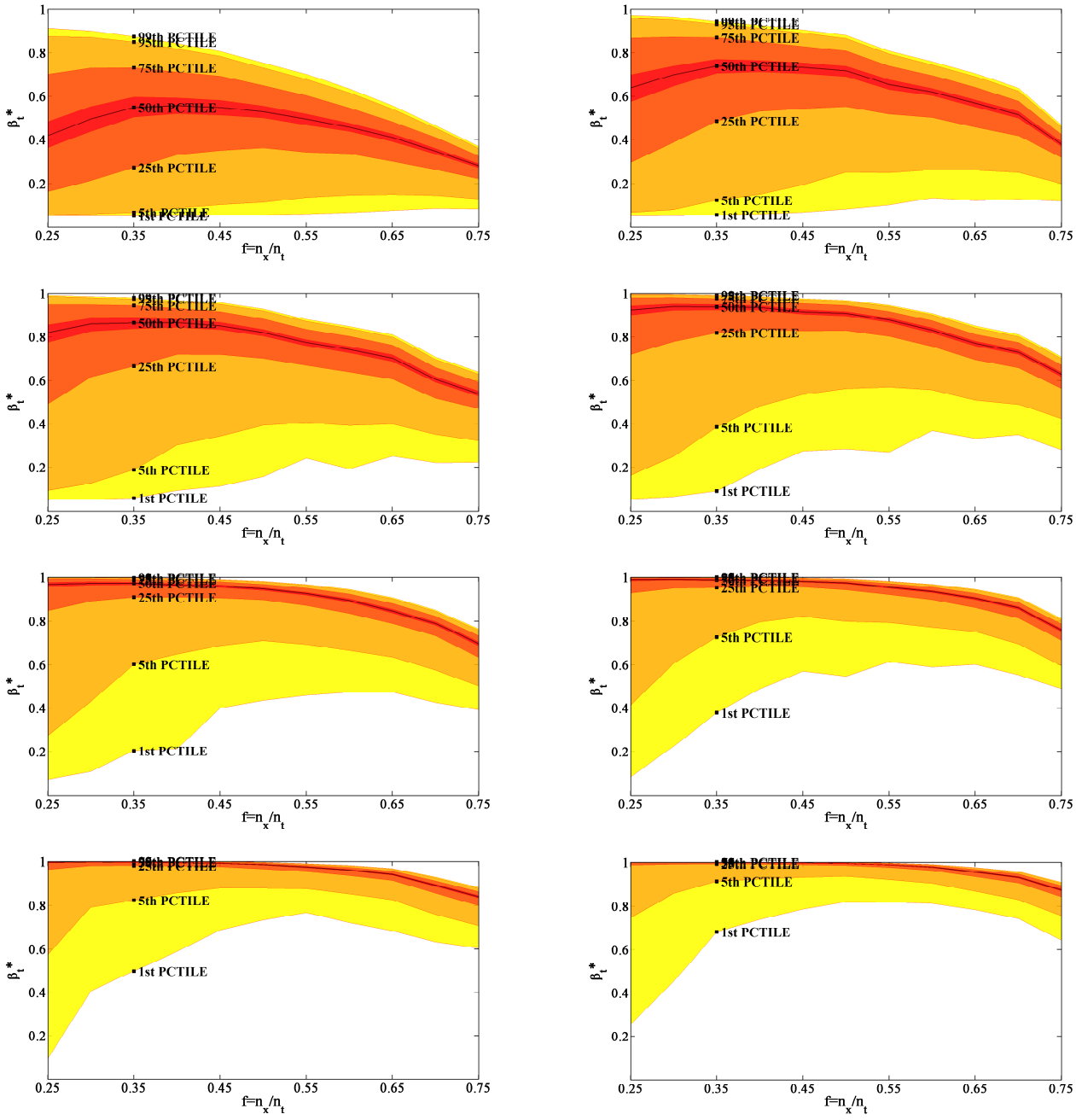



Figure 5: Simulation-based approximation of selected percentiles of the distributions of the power function $\beta_t^\star$ plotted against the sample allocation for total sample size $n_T = 20, 25, 30, 35, 40, 45, 50, 55$. The parameter values used here are: $\theta = (0.3, 0.3, 0.5, 0.5)'$, $\Sigma$ compound symmetric with $\sigma = 0.6$ ,$\rho = 0.6$. Jeffrey's priors are used to derive Bayesian posterior estimates. Higher power levels are achieved for $f \in [0.3, 0.5]$. As $n_T$ increases, the power distributions are more tight and they attain higher levels.

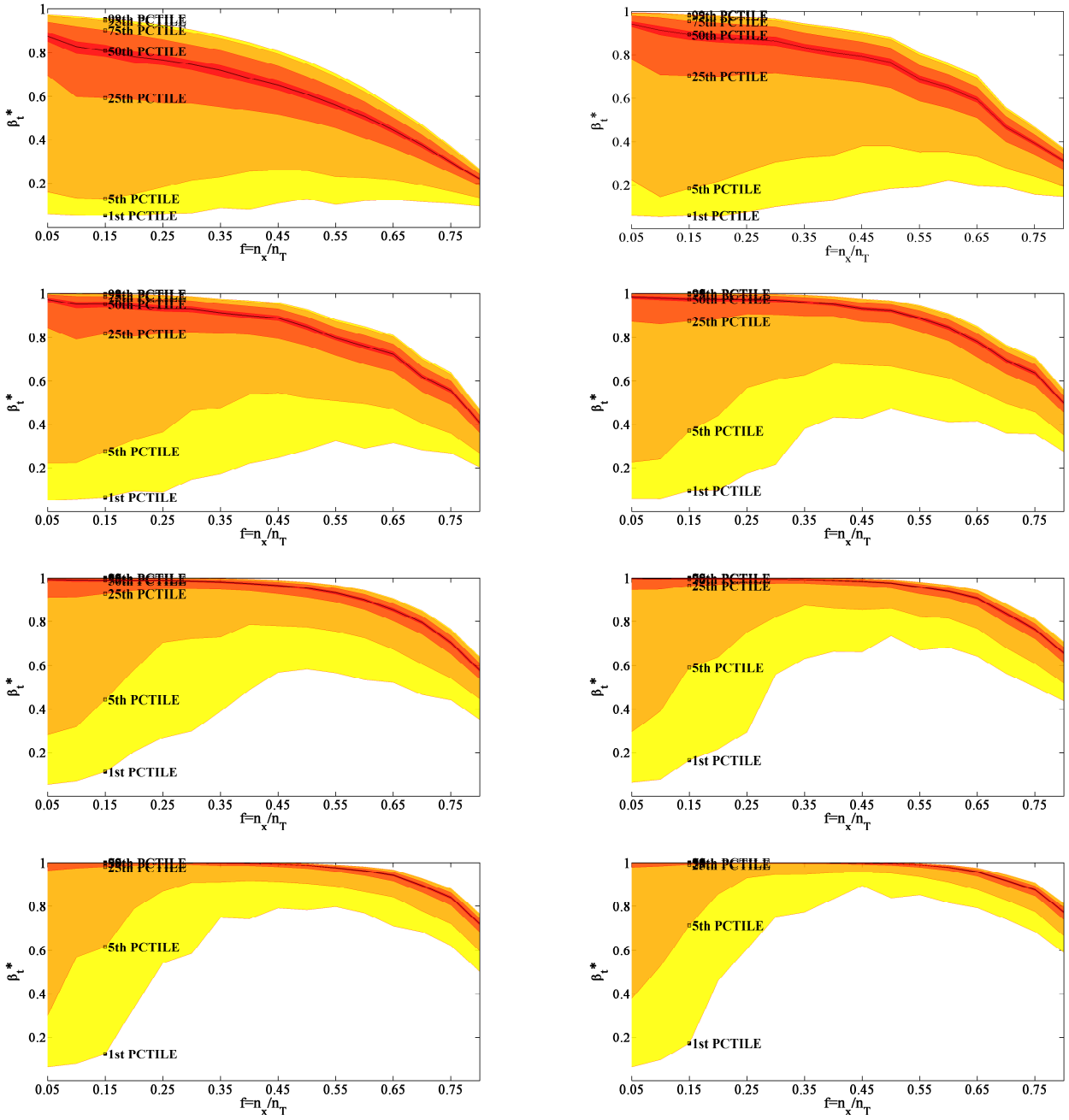



Figure 6: Simulation-based approximation of selected percentiles of the distributions of the power function $\beta_t^\star$ plotted against the sample allocation for total sample size $n_T = 20, 25, 30, 35, 40, 45, 50, 55$. The parameter values used here are: $\theta = (0.3, 0.3, 0.5, 0.5)'$, $\Sigma$ compound symmetric with $\sigma = 0.6$ ,$\rho = 0.6$. Conjugate priors with hyperparameters $n_0 = 1$, $\theta_0 = (0.5, 0.5, 0.5, 0.5)'$, $\nu_0 = 1$, $S_0$ compound symmetric with $s_0^2 = 1$ and $r_0 = 0.7$ are used to derive Bayesian posterior estimates. The upper percentiles of $\beta_t^\star$ take higher levels for $f = 0.05$ while lower percentiles attain higher levels for $f \in (0.3, 0.5)$. As $n_T$ increases, the power distributions are more tight and they attain higher levels.

## I.2 SS and PC tests

Läuter et al. in [26] propose a method in which the weighting vectors of the linear combinations are selected using the observed data, while controlling the false positive rate. Läuter et al. prove that if the weighting vector is completely determined by the sums of product matrix $Y'Y$, where $Y = (Y_{ij})_{i=1,k=1}^{i=n_y,k=K}$ then the linear combination $t$-statistic is Student-$t$ distributed with $n_y - 1$ degrees of freedom. Two forms of weighting vectors are proposed, the first of which is equal to the diagonal of $Y'Y$, i.e. $w_k^{SS} = 1/\sqrt{\left(\sum_{i=1}^{n_y} y_{ik}^2\right)}$, $k = 1, ..., K$. Note that $\frac{1}{n_y}\sum_{i=1}^{n_y} y_{ik}^2$ is the unbiased estimator of $E\left(Y_k^2\right) = Var\left(Y_k\right) + \left(E\left(Y_k\right)\right)^2$. Therefore, correlations between the components of

the response are not taken into account and components with larger expectation are expected to be down-weighed which contradicts the intuition for selecting weights. On the other hand, components with larger variance are down-weighed which intuitively is a desirable property. The second form of weighting vector proposed in [26] is $w^{PC} = Diag(Y'Y)^{-1/2}\overline{w}$, where $\overline{w}$ is the standarised ($\overline{w}'\overline{w} = 1$) eigenvector corresponding to the largest eigenvalue of the matrix $Diag(Y'Y)^{-1/2}(Y'Y)Diag(Y'Y)^{-1/2}$ and $Diag(Y'Y)$ is a diagonal matrix with the same diagonal with $Y'Y$. Note that the two proposed weighting vectors are proportional, i.e. $w_k^{PC} = \overline{w}_k w_k^{SS}$, $k = 1, 2, ..., K$, which indicates that under certain conditions they have similar behaviour. Läuter et al. suggest that the weighting vectors, $w^{SS}, w^{PC}$ and the corresponding linear combination $t$-tests, $t^{SS}$ and $t^{PC}$, respectively, are powerful under the one-factor model while Frick [11] shows that $t^{SS}$ and $t^{PC}$ lack power when $\mu$ have at least one zero entry, even for extremely large sample sizes and/or magnitudes of $\mu$. In the results presented in section 4.3, these tests lack power when the components of the vector of $\mu$ do not have the same or similar size.

# References


[1] James Ambrose. Computerized transverse axial scanning (tomography): Part 2. clinical application. *Br J Radiol*, 46(552):1023–1047, December 1973.

[2] T.W. Anderson. *An introduction to multivariate statistical analysis, 2nd edition.* John Wiley and Sons, 2003.

[3] S. Y. Bookheimer, M. H. Strojwas, M. S. Cohen, A. M. Saunders, M. A. Pericak-Vance, J. C. Mazziotta, and G. W. Small. Patterns of brain activation in people at risk for alzheimer's disease. *New England Journal of Medicine*, 343:450 – 456, 2000.

[4] H. C. Breiter, R. L. Gollub, R. M. Weisskoff, D. N. Kennedy, N. Makris, J. D. Berke, J. M. Goodman, H. L. Kantor, D. R. Gastfriend, J. P. Riorden, R. T. Mathew, B. R. Rosen, and Hyman S. E. Acute effects of cocaine on human brain activity and emotion. *Neuron*, 19:591–611, 1997.

[5] Christine Brideau, Bert Gunter, Bill Pikounis, and Andy Liaw. Improved statistical methods for hit selection in high-throughput screening. *Journal of Biomolecular Screening*, 8(6):634–647, December 2003.

[6] Ralph B. D'agostino and Heidy K. Russell. *Multiple Endpoints, Multivariate Global Tests.* John Wiley & Sons, Ltd, 2005.

[7] R. J. Davidson, W. Irwin, M. J. Anderle, and N. H. Kalin. The Neural Substrates of Affective Processing in Depressed Patients Treated With Venlafaxine. *Am J Psychiatry*, 160(1):64–75, 2003.

[8] A. Dmitrienko, A. C. Tamhane, and F. Bretz. *Multiple Testing Problems in Pharmaceutical Statistics.* Chapman and Hall/CRC, 2010.

[9] Alex Dmitrienko, Ajit C. Tamhane, and Frank Bretz, editors. *Multiple Testing Problems in Pharmaceutical Statistics (Chapman & Hall/Crc Biostatistics Series.* Chapman & Hall/CRC, 2009.

[10] William E. Evans and Mary V. Relling. Pharmacogenomics: Translating functional genomics into rational therapeutics. *Science*, 286(5439):487–491, October 1999.

[11] H. Frick. On the power behaviour of lauter's exact multivariate one-sided tests. *Biometrical Journal*, 38:405–414, 1996.

[12] K. Friston, J. Ashburner, S. Kiebel, T. Nichols, and W. Penny. *Statistical parametric mapping: the analysis of funtional brain images.* Elsevier/Academic Press, 2007.

[13] A. Gelman, J. B. Carlin, H. S. Stern, and D. B. Rubin. *Bayesian Data Analysis.* Chapman and Hall/CRC, 2004.

[14] TR Golub, DK Slonim, P Tamayo, C Huard, M Gaasenbeek, JP Mesirov, H Coller, ML Loh, JR Downing, MA Caligiuri, CD Bloomfield, and ES Lander. Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science*, 286(5439):531–7–, October 1999.

[15] M. Hämäläinen, R. Hari, R. J. Ilmoniemi, J. Knuutila, and O. V. Lounasmaa. Magnetoencephalography-theory, instrumentation, and applications to noninvasive studies of the working human brain. *Reviews of Modern Physics*, 65:413–497, April 1993.

[16] G. D. Honey, E. T. Bullmore, W. Soni, M. Varatheesan, S. C. R. Williams, and T. Sharma. Differences in frontal cortical activation by a working memory task after substitution of risperidone for typical antipsychotic drugs in patients with schizophrenia. *Proceedings of the National Academy of Sciences of the United States of America*, 96(23):13432–13437, 1999.

[17] Garry Honey and Ed Bullmore. Human pharmacological mri. *Trends in Pharmacological Sciences*, 25(7):366 – 374, 2004.

[18] Harold Hotelling. The generalization of student's ratio. *The Annals of Mathematical Statistics*, 2(3), 1931.

[19] S. A. Huettel, A. W. Song, and G. McCarthy. *Functional Magnetic Resonance Imaging*. Sinauer Associates, 2004.

[20] L.W. Huson. The Bayesian Bootstrap in a predictive power analysis. *Case Studies in Business, Industry and Government Statistics*, 3:18–22, 2009.

[21] G. D. Iannetti, L. Zambreanu, R. G. Wise, T. J. Buchanan, J. P. Huggins, T. S. Smart, W. Vennart, and I. Tracey. Pharmacological modulation of pain-related brain activity during normal and central sensitization states in humans. *Proceedings of the National Academy of Sciences of the United States of America*, 102(50):18195–18200, 2005.

[22] ICH. International harmonised tripartite guideline: Statistical principles for clinical trials e9. http://www.ich.org/LOB/media/MEDIA485.pdf, February 5, 1998.

[23] P. Jezzard, P. M. Matthews, and S. M. Smith. *Functional MRI: An Introduction to Methods*. Oxford University Press, 2001.

[24] Peter K. Kimani, Nigel Stallard, and Jane L. Hutton. Dose selection in seamless phase ii/iii clinical trials based on efficacy and safety. *Statist. Med.*, 28(6):917–936, 2009.

[25] Gillian A. Lancaster, Susanna Dodd, and Paula R. Williamson. Design and analysis of pilot studies: recommendations for good practice. *Journal of Evaluation in Clinical Practice*, 10(2):307–312, 2004.

[26] J. Lauter, E. Glimm, and S. Kroph. New multivariate tests for data with an inherent structure. *Biometrical Journal*, 38:1–23, 1996.

[27] E. L. Lehmann and J. P. Romano. *Testing statistical hypotheses*. Springer, 2005.

[28] B. R. Logan and A. C. Tamhane. On o'brien's ols and gls tests for multiple endpoints. *Lecture Notes-Monograph Series*, 47:76–88, 2004.

[29] Ricardo Macarron, Martyn N. Banks, Dejan Bojanic, David J. Burns, Dragan A. Cirovic, Tina Garyantes, Darren V. S. Green, Robert P. Hertzberg, William P. Janzen, Jeff W. Paslay, Ulrich Schopfer, and G. Sitta Sittampalam. Impact of high-throughput screening in biomedical research. *Nat Rev Drug Discov*, 10(3):188–195, March 2011.

[30] Nathalie Malo, James A Hanley, Sonia Cerquozzi, Jerry Pelletier, and Robert Nadon. Statistical practice in high-throughput screening data analysis. *Nat Biotech*, 24(2):167–175, February 2006.

[31] K. V. Mardia, J.T. Kent, and J.M. Bibby. *Multivariate analysis*. Academic Press Inc., 1979.

[32] Paul M. Matthews, Garry D. Honey, and Edward T. Bullmore. Applications of fmri in translational medicine and clinical practice. *Nat Rev Neurosci*, 7(9):732–744, September 2006.

[33] G. D. Mitsis, G. D. Iannetti, T. S. Smart, I. Tracey, and R. G. Wise. Regions of interest analysis in pharmacological fMRI: How do the definition criteria influence the inferred result? *Neuroimage*, 40:121–132, 2007.

[34] M. Neuhauser. How to deal with multiple endpoints in clinical trials. *Fundamental and Clinical Pharmacology*, 20:515–523, 2006.

[35] P. C. O'Brien. Procedures for comparing samples with multiple endpoints. *Biometrics*, 40:1079–1087, 1984.

[36] A. O'Hagan and J. W. Stevens. Bayesian assessment of sample size for clinical trials of cost-effectiveness. *Medical Decision Making*, 21, 2001.

[37] Homer H. Pien, Alan J. Fischman, James H. Thrall, and A.Gregory Sorensen. Using imaging biomarkers to accelerate drug development and clinical trials. *Drug Discovery Today*, 10(4):259–266, February 2005.

[38] S. J. Pocock, N. L. Geller, and A. A. Tsiatis. The analysis of multiple endpoints in clinical trials. *Biometrics*, 43, 1987.

[39] R. A. Poldrack. Region of interest analysis for fmri. *Social Cognitive and Affective Neuroscience*, 2:67–70, 2007.

[40] Russell A. Poldrack and Jeanette A. Mumford. Independence in roi analysis: where is the voodoo? *Social Cognitive and Affective Neuroscience*, 4(2):208–213, June 2009.

[41] Russell A. Poldrack, Jeanette A. Mumford, and Thomas E. Nichols. *Handbook of functional MRI data analysis*. Cambridge University Press, Cambridge, 2011.

[42] A. J. Sankoh, R. B. D'Agostino, and M. F. Huque. Efficacy endpoint selection and multiplicity adjustment methods in clinical trials with inherent multiple endpoint issues. *Statistics in Medicine*, 22:3133–3150, 2003.

[43] Jay Shendure and Hanlee Ji. Next-generation dna sequencing. *Nat Biotech*, 26(10):1135–1145, October 2008.

[44] Weichung Joseph Shih, Pamela A. Ohman-Strickland, and Yong Lin. Analysis of pilot and early phase studies with small sample sizes. *Statist. Med.*, 23(12):1827–1842, 2004.

[45] Richard Simon and Peter F. Thall. *Phase II Trials*. John Wiley & Sons, Ltd, 2005.

[46] D. J. Spiegelhalter, K. R. Abrams, and J. P. Myles. *Bayesian approaches to clinical trials and health-care evaluation*. John Wiley and Sons, 2004.

[47] D. J. Spiegelhalter and L. S. Freedman. A predictive approach to selecting the size of a clinical trial. *Statistics in Medicine*, 5:1–13, 1986.

[48] D. J. Spiegelhalter, L. S. Freedman, and P. Blackburn. Monitoring clinical trials: conditional or predictive power? *Controlled Clinical Trials*, 7:8–17, 1986.

[49] Ludwig G. Strauss and Peter S. Conti. The applications of pet in clinical oncology. *J Nucl Med*, 32(4):623–648, April 1991.

[50] B. C. Sutradhar. On the characteristic function of multivariate Student t-distribution. *The Canadian Journal of Statistics*, 14:329–337, 1986.

[51] D. Tang, N. L. Geller, and S. J. Pocock. On the design and analysis of randomized clinical trials with multiple endpoints. *Journal of the American Statistical Association*, 84:776–779, 1993.

[52] G. Wassmera, C. P. Reitmeirb, M. Kieserc, and W. Lehmachera. Procedures for testing multiple endpoints in clinical trials: An overview. *Journal of Statistical Planning and Inference*, 82:69–81, 1986.

[53] Peter H. Westfall, Alok Krishen, and S. Stanley Young. Using prior information to allocate significance levels for multiple endpoints. *Statist. Med.*, 17(18):2107–2119, 1998.

[54] B. Whitcher and P. Matthews. Noninvasive brain imaging for Experimental medicine in drug discovery and development: Promise and pitfalls. *International Journal of Pharmaceutical Medicine*, 20:167–175(9), 2006.

[55] R. G. Wise and I. Tracey. The role of fMRI in drug discovery. *Journal Of Magnetic Resonance Imaging*, 23:862–876, 2006.