Published in final edited form as:
J Am Stat Assoc. 2014 July ; 109(507): 967–976. doi:10.1080/01621459.2014.922886.

Causal Inference for fMRI Time Series Data with Systematic Errors of Measurement in a Balanced On/Off Study of Social Evaluative Threat

Michael E. Sobel[*] and
Department of Statistics Columbia Unversity

Martin A. Lindquist
Department of Biostatistics Johns Hopkins University

Abstract

Functional magnetic resonance imaging (fMRI) has facilitated major advances in understanding human brain function. Neuroscientists are interested in using fMRI to study the effects of external stimuli on brain activity and causal relationships among brain regions, but have not stated what is meant by causation or defined the effects they purport to estimate. Building on Rubin's causal model, we construct a framework for causal inference using blood oxygenation level dependent (BOLD) fMRI time series data. In the usual statistical literature on causal inference, potential outcomes, assumed to be measured without systematic error, are used to define unit and average causal effects. However, in general the potential BOLD responses are measured with stimulus dependent systematic error. Thus we define unit and average causal effects that are free of systematic error. In contrast to the usual case of a randomized experiment where adjustment for intermediate outcomes leads to biased estimates of treatment effects (Rosenbaum, 1984), here the failure to adjust for task dependent systematic error leads to biased estimates. We therefore adjust for systematic error using measured “noise covariates” , using a linear mixed model to estimate the effects and the systematic error. Our results are important for neuroscientists, who typically do not adjust for systematic error. They should also prove useful to researchers in other areas where responses are measured with error and in fields where large amounts of data are collected on relatively few subjects. To illustrate our approach, we re-analyze data from a social evaluative threat task, comparing the findings with results that ignore systematic error.

Keywords

Balanced design; BOLD contrast; Causal inference; fMRI; Longitudinal data; Measurement error

1 Introduction

Recent developments in neuroimaging have facilitated advances in the understanding of human brain function, the emergence of new fields such as cognitive neuroscience and neuroeconomics, and the revitalization of more established fields such as psychophysiology

[*] Corresponding author: 1255 Amsterdam Avenue, Columbia University, New York, NY 10027, michael@stat.columbia.edu..

and social psychology. Dating back to the early 1990's, functional magnetic resonance imaging (fMRI), a safe and non-invasive imaging technology with reasonable spatial and temporal resolution, has become the primary imaging modality employed in these areas.

Neuroscientists using fMRI are increasingly interested in making causal inferences about brain function (Ramsey et al., 2010; Valdes-Sosa et al., 2011). At one end of the scale, investigators want to estimate the effects of stimuli on activity in specific brain locations, for example, determining which regions of the brain are involved in processing physical pain (Tracey, 2008). At the other end, investigators interested in “effective connectivity” wish to study directed influences among brain regions, for example, how attention modulates directional connectivity between visual regions V1 and V5 (Friston et al., 2003). Various methods, including Granger causal mapping (Roebroeck et al., 2005), dynamic causal models (Friston et al., 2003), structural equation models (McIntosh and Gonzalez-Lima, 1994), directed graphical models (Ramsey et al., 2010), even methods borrowed from network analysis (Bassett and Bullmore, 2006), have been used. However, neuroscientists using these procedures have not stated what they mean by causation or defined the effects they purport to be estimating. Unless these issues are adequately addressed, it will not be possible to either assess whether (or under what conditions) such procedures can be used to make valid causal inferences or develop appropriate methodologies and experimental designs for assessing causal relations in brain research.

The approach herein builds on Rubin's causal model and its extension to longitudinal data. We use potential outcomes $Y_{it}(z_1, ..., z_t)$ to indicate the response subject i would have in period t to a sequence of treatments $Z_1, ..., Z_t$, with Z_t set to z_t, $t = 1, ..., T$; for subject i, the unit effect of sequence $z_1, ..., z_t$ versus $z^*_1 , ..., z^*_t$ might then be defined as $Y_{it}(z_1, ..., z_t) - Y_{it}(z^*_1 , ..., z^*_t)$. Average effects are then obtained by averaging over subjects. Our approach should not be confused with Granger causality, which is sometimes used to model effective connectivity. There, a time series $Z_1, ...Z_t$ would be said to Granger cause Y_{t+1} if, when added to $Y_1, ..., Y_t, Z_1, ..., Z_t$ accounted for additional variation in Y_{t+1}; thus, causality is defined operationally and equated with association, without regard for potentially different values of the response under different treatment sequences.

In fMRI experiments, a multivariate time series of three dimensional digital images of the brain is obtained for each subject, each image composed of thousands of equally sized volume elements (voxels). At each voxel, the blood oxygenation level dependent (BOLD) contrast is measured. The resulting data are then most frequently analyzed using the so called “massively univariate approach” (Lindquist, 2008). For each subject, the relationship between the sequence of stimuli and the BOLD responses is analyzed voxel by voxel. In “group analyses”, the results from each voxel are then combined across subjects in a second stage. To determine whether or not a voxel is activated, a threshold that takes into account the multiple testing problem is chosen. As active voxels are spatially grouped, cluster level inference (Poldrack et al., 2011), in which the preceding analysis is used in conjunction with the information on voxel positions to infer activation in empirically constructed clusters of voxels, is typically conducted. In contrast, in region of interest analyses, where interest centers on activity in predefined brain regions consisting of hundreds or thousands of voxels, researchers will typically average the voxel values in the region and analyze the

relationship between the stimuli and these averages. In connectivity analyses, where the association between brain activity in different regions is studied, researchers will sometimes average over subjects as well.

Regardless of the ultimate scale of focus, we believe that a principled framework for causal inference in functional neuroimaging should start with the most elemental outcome, the BOLD response of a single subject at a particular voxel and time; if desired, functions of the elemental responses can always be used to define outcomes at higher levels of organization, for example, regions. Potential outcomes notation is used to represent these responses and formalize the idea that causal relationships sustain counterfactual conditional statements. Though widely used in the statistical literature on causal inference, this notation has only recently been introduced into the neuroscience literature (Lindquist and Sobel, 2011, 2013; Lindquist, 2012; Luo et al., 2012).

As in the statistical literature on longitudinal causal inference (see Robins and Hernán (2009) for a nice overview), subjects are treated as the experimental units, with each unit $i = 1, ..., n$ observed for $t = 1, ..., T$ periods on one or more occasions. Our framework, however, is not merely a routine application of ideas from this literature. There potential outcomes are used to define unit and average causal effects. As only one sequence per subject is observed, the unit effects are treated as unidentified. Average effects are identified under assumptions such as sequential ignorability and positivity (in addition to the stable unit treatment value assumption). In the simplest case, if subjects are randomly assigned at baseline to one of R treatment regimens of interest, the assignment mechanism is strongly ignorable (Rosenbaum and Rubin, 1983) and the difference in expectation at time t between subjects assigned to regimens A and B is also the effect of assignment to regimen A vs. B at time t, as the T dimensional vectors of treatment assignments and potential outcomes are independent. If also R is “small” relative to n, enough subjects can be assigned to each regimen so that estimation can proceed non-parametrically, comparing the sample means in period t of the subjects in each group.

However, the BOLD response is measured with systematic error, generally task dependent; thus, unit and average effects defined in the manner above will evidence both causation and systematic error. Therefore, we define unit (and average) effects free of systematic error. In the simplest case above, it is well known (Rosenbaum, 1984) that adjusting for intermediate outcomes affected by the cause leads to biased estimates of average treatment effects; here, however, failing to adjust for stimulus dependent systematic error leads to biased estimates. Our results are important for neuroscientists, who typically do not adjust estimates of treatment effects for systematic measurement error. They should also prove useful to researchers in other areas where responses are measured with error, for example, the social and behavioral sciences.

In addition, n is often “small” relative to R, in which case the majority of regimens of interest are not observed. For example, Robins et al. (1999) estimate effects of maternal stress on children's health in an observational study with $T = 30$, $R = 2^{30}$ possible regimens, and $n = 167$ mother-child pairs. Although the identification conditions above may hold, estimates comparing regimens where one or both regimens are unobserved (which constitute

the vast majority of comparisons) necessarily rely on extrapolation from the model fitted to the observed data, not on direct comparisons between subjects exposed to different regimes.

The small *n*, large *R* case is also common in functional neuroimaging, where $T > 100$ and $n < 30$ are typical. In this case, even if the experimenter were to randomly assign subjects to each regimen of interest with positive probability, thereby satisfying the identification conditions above, the vast majority of regimens will be unobserved and, as above, estimated comparisons between regimens will be based primarily on extrapolations from a model for the observed data. Further, in actual neuroimaging experiments, often only a small fraction of the *R* regimens of interest are assigned a positive probability of observation, and in one of the most widely used experimental designs in functional neuroimaging, the so-called "balanced design", the experimenter simply chooses a regimen to which all subjects are assigned. However, unlike the case of an observational study, where a regimen with 0 probability is one that would not be observed in the population from which the sample is drawn and may therefore not be of interest, in neuroimaging experiments, balanced or otherwise, the experimenter could have assigned subjects to such a regimen with non-zero probability. In this case, as in an observational study where the *R* regimens satisfy the positivity assumption, extrapolation from the model for the observed data to unobserved regimens of interest will be warranted under suitable ignorability conditions and/or if model parameters governing treatment effects are invariant across regimens.

The longitudinal causal inference literature has focused on the estimation of average effects, using marginal or nested structural models. Estimation in observational studies or sequentially randomized experiments typically requires modeling and taking into account the assignment mechanism in each period, which may depend on previous treatments, responses and covariates. In functional neuroimaging experiments, where subjects are typically assigned at random to treatment regimens or all subjects are assigned to a particular regimen, estimation is simplified, as it is not necessary to also model the assignment mechanism. To model the BOLD response, a number of modeling assumptions specific to the field of functional neuroimaging are used, in conjunction with a linear mixed model, to estimate both subject specific (unit) and average causal effects, and the variance of the unit effects. We illustrate our approach with data from a balanced social evaluative threat (SET) experiment to study the effect of the task on activation in different brain locations.

The vast majority of research in functional neuorimaging uses the so-called General Linear Model (GLM) approach (Friston et al., 1995), in which each fMRI time series is modeled as a linear combination of several different signal components, to estimate treatment effects. However, the estimates may depend upon the components included, which are typically chosen in an ad-hoc manner. For example, when subjects move their heads in the MRI scanner, the sequence of voxel values corresponding to a given voxel *v* in the resulting images may actually be composed of values from different brain locations, necessitating motion correction, in which, prior to analysis, researchers estimate the between scan movement using a rigid body transformation, and then realign the images. But this does not correct for changes in the magnetic field caused by head motion that lead to nonlinear, time-varying distortion of the resulting brain images, and there has been some debate on how to deal with these residual artifacts (Johnstone et al., 2006). Some researchers (Johnstone et al.,

2006; Lund et al., 2006) suggest including vectors of motion regressors as "nuisance covariates" in the model for the BOLD response to adjust for this error, arguing that this yields estimates that seem more reasonable than those obtained when these covariates are omitted. But as head motion tends to be task related, there is concern that inclusion of these covariates can lead to underestimating the signal component due to "genuine" activiation (Poldrack et al., 2011). For example, Churchill et al. (2012) argue that inclusion of motion regressors has a "generally detrimental effect on activation" for subjects with minimal head motion.

We do not see how the issue of whether or not to include motion regressors can be properly resolved when the effects researchers purport to estimate are not even defined. Our framework, in which the treatment effects are explicitly defined, allows resolution of this issue in a principled fashion, yielding the conclusion that omitting signal components due to task dependent systematic error will generally lead to biased estimates of effects. Thus, motion regressors should be included in models for the BOLD response. This is an important conclusion, especially as a recent survey of the neuroimaging literature (Carp, 2012) reports that less than one third of studies (31.5%) included motion regressors in models for the BOLD response.

We proceed as follows. Section 2 describes the SET task data. Section 3 discusses the sources of systematic error in BOLD responses, and puts forth a framework for causal inference in longitudinal fMRI experiments. In Section 4, we apply the framework to the SET task data, finding activation in the visual cortex and superior temporal cortices during the early part of the speech preparation, and activation in the ventromedial prefrontal cortex throughout the entire preparation. We also illustrate the bias that occurs when motion regressors are excluded, finding significant activation in the ventricles and at the edge of the brain, symptomatic of motion-related artifacts. Section 5 concludes, discussing extensions of the framework to brain regions (collections of voxels) and functional mediation.

2 The SET Task Data Set

The data, a time series of 215 images acquired every two seconds for each of 25 subjects, were collected to study the effects of an anxiety producing task on brain activity and heart rate (Wager et al., 2009). An off/on/off balanced design was used. Subjects were initially told they would be asked to prepare a 7 minute speech for possible presentation to their peers. At the start of scanning, they viewed a fixation cross for 2 minutes (resting baseline). They then viewed a slide for 15 seconds describing the topic (why subject is a good friend) and instructing subjects to prepare enough material for the entire 7 minutes. After 2 minutes of silent preparation, a 15 second relief instruction, telling subjects they would not have to give the speech, was given. Subjects were then scanned for an additional 2.5 minutes. More details regarding data acquisition can be found in Section 4.

Single subject analyses of fMRI data assume the voxel values in an image are simultaneously recorded and the brain location corresponding to a given voxel v in the sequence of 215 images is fixed. In analyses with multiple subjects, it is also assumed voxel v corresponds to the same brain location for each of the subjects. However, each 3

dimensional image consists of two dimensional slices acquired at different times during the 2 second interval. Second, even under the best conditions, there will be some head motion during scanning. Third, human brains vary in size and shape. Thus, to render these assumptions valid, neuroscientists "pre-process" the data. Specifically, a slice timing correction was employed to adjust for the different acquisition times of the two dimensional slices. The images were then corrected for "bulk" motion and each subjects' data registered to a subject-specific structural image in a process known as co-registration. The co-registered images from different subjects were then spatially normalized to a standard template. Finally, the images were spatially smoothed using a Gaussian kernel. The BOLD responses in the resulting images are the outcomes we analyze. An important point, addressed in the next section, is that even after preprocessing, these responses contain both systematic (generally stimulus dependent) and random errors of measurement. For a detailed description of the preprocessing performed in this experiment see Section 4; for more information about fMRI pre-processing in general see Lindquist (2008).

3 Causal Inference for Longitudinal fMRI Data with Systematic Errors in Outcomes

3.1 Assumptions

Subjects $i = 1, ..., n$ are observed, starting at subject specific times $k_i + 1$, for $t = 1, .., T$ equally spaced periods. In each period, either no stimulus or one of $j \in \{1, ..., J\}$ stimuli is administered. Define $z_{jt} = 1$ if stimulus j is applied in period t, 0 otherwise, $j = 1, ..., J$, $z_t = (z_{1t}, ..., z_{Jt})$ the assignment vector at time t, and $\bar{z_T} \equiv (z_1, ..., z_T)$ the treatment regimen. In principle any of the $(J + 1)^T$ possible treatment regimens can be administered to a subject; in practice only a subset Ω of these may be of interest. The notation $\bar{z_t} \equiv (z_1, ...,z_t)$ is used to denote a sub-regimen of $\bar{z_T}$ through period t. For each subject i and voxel $v \in \{1, ..., V\}$, we consider the potential BOLD series $Y_{iv,ki+1}(\bar{z_T})$, ..., $Y_{iv,ki+T}(\bar{z_T})$ for each $\bar{z_T} \in \Omega$. Throughout we make the stable unit treatment value assumption (SUTVA) that a subject's outcomes do not depend on the regimens to which other subjects are exposed (Rubin, 1980); though implausible in some contexts, this assumption is justifiable here.

In the previous section, we described preprocessing steps taken to render valid assumptions about the spatial and temporal alignment of the BOLD responses. However, even after such steps, the BOLD responses contain both random error and systematic error due to scanner drift and head motion not corrected for during preprocessing. We decompose the potential responses as:

(A1) BOLD Response Decomposition — For all $\bar{z_T} \in \Omega$,

$$Y_{iv,k_i+t}\left(\bar{\mathbf{z}}_T\right) = \Psi_{iv,k_i+t}\left(\bar{\mathbf{z}}_T\right) + B_{iv,k_i+t}\left(\bar{\mathbf{z}}_T\right) + \varepsilon_{iv,k_i+t}\left(\bar{\mathbf{z}}_T\right), \quad (1)$$

where $\Psi_{iv,ki+t}(\bar{z_T})$ is the subject's "true" BOLD response (signal), $\varepsilon_{iv,ki+t}(\bar{z_T})$ is a random error with mean 0, and $B_{iv,ki+t}(\bar{z_T})$ is the systematic error or measurement bias for subject i at voxel v at time $k_i + t$. Note that here the period t response may depend on treatments administered in subsequent periods; assumption (A3) below addresses this issue further.

In equation (1), $\Psi_{iv,ki+t}(z_T^-)$ is allowed to depend on the arbitrary starting time k_i. As in classical psychometric test theory (Lord and Novick, 1968), we regard the "true" response as a stable latent trait of the individual, leading to the defining constraint:

(A2) True Response Time Invariance — $\forall i$, $z_T^- \in \Omega$, and (k_i, k'_i), $\Psi_{iv,ki+t}(z_T^-) = \Psi_{iv,k'i+t}(z_T^-) \equiv \Psi_{ivt}(z_T^-)$.

Similarly to the classical theory, $\Psi_{ivt}(z_T^-)$ may be defined as the expectation of $\Psi_{iv,ki+t}(z_T^-)$ over repeated occasions: $\Psi_{iv,ki+t}(z_T^-) \equiv \Psi_{ivt}(z_T^-) + \varepsilon_{iv,ki+t}(z_T^-)$, where $\varepsilon_{iv,ki+t}(z_T^-)$ has mean 0; $\varepsilon_{iv,ki+t}(z_T^-)$ can then be absorbed into the error $\varepsilon_{iv,ki+t}(z_T^-)$.

We now define the unit effects. For subject i, the effect of regime z_T^- vs. $z^{*}{}_T^-$ in period $t = 1, ..., T$ would normally be defined as $Y_{iv,ki+t}(z_T^-) - Y_{iv,ki+t}(z^{*}{}_T^-)$. But if the systematic error is task dependent (for example, task related head-motion not removed during pre-processing), such a definition incorporates errors $B_{iv,ki+t}(z_T^-) - B_{iv,ki+t}(z^{*}{}_T^-)$ as a component of the unit effect. Second, even in the absence of systematic error, the observed BOLD response is subject to random errors of measurement produced by cardiac and respiratory activity, for example. Thus, for subjects $i = 1, ..., n$, the unit effects of treatment regimen z_T^- vs. $z^{*}{}_T^-$ in period $t = 1, ..., T$ are defined using the "true" responses:

$$\psi_{ivt}\left(\bar{\mathbf{z}}_T, \bar{\mathbf{z}}^{*}_T\right) \equiv \mathbf{\Psi}_{ivt}\left(\bar{\mathbf{z}}_T\right) - \mathbf{\Psi}_{ivt}\left(\bar{\mathbf{z}}^{*}_T\right). \quad (2)$$

Note that as a consequence of (A2), the unit effects do not depend on the start times k_i; however, the BOLD responses $Y_{iv,ki+t}(z_T^-)$ may depend on k_i.

For subject i, the effect $\phi_{ivt}(z_T^-, z^{*}{}_T^-)$ is treated as a parameter. When subjects are sampled from a population of interest, $\phi_{ivt}(z_T^-, z^{*}{}_T^-)$ is a random variable and the average treatment effect

$$E\left(\psi_{ivt}\left(\bar{\mathbf{z}}_T, \bar{\mathbf{z}}^{*}_T\right)\right), \quad (3)$$

and the variance of subject effects

$$V\left(\psi_{ivt}\left(\bar{\mathbf{z}}_T, \bar{\mathbf{z}}^{*}_T\right)\right) = E\left(\psi_{ivt}\left(\bar{\mathbf{z}}_T, \bar{\mathbf{z}}^{*}_T\right) - E\left(\psi_{ivt}\left(\bar{\mathbf{z}}_T, \bar{\mathbf{z}}^{*}_T\right)\right)\right)^2, \quad (4)$$

where the expectation is now taken over subjects, are typically of interest. Next, we assume:

(A3) Temporal Consistency — For all $z_T^- \equiv \Omega$, the BOLD response $Y_{iv,ki+t}(z_T^-)$ in period t and its components $_{ivt}(z_T^-)$, $B_{iv,ki+t}(z_T^-)$, $\varepsilon_{iv,ki+t}(z_T^-)$ do not depend on stimuli administered after period t.

Thus, for regimen $z_T^- \equiv (z_t^-, z_{t+1}, ..., z_T)$, we may write $Y_{iv,ki+t}(z_T^-) = Y_{iv,ki+t}(z_t^-)$, $\Psi_{ivt}(z_T^-) = \Psi_{ivt}(z_t^-)$, $\phi_{ivt}(z_T^-) = \phi_{ivt}(z_t^-)$, etc. Assumption (A3) is so ubiquitous in neuroimaging studies and elsewhere that it is rarely made explicit. Under this assumption, the responses are not related to future stimuli, as might occur if stimuli are administered in a predictable order at

fixed time intervals. In practice, neuroimaging researchers circumvent this problem by avoiding regimens that appear to follow a predictable pattern; such regimens lie in the complement of Ω. In randomized studies where multiple regimens are observed, assumption (A3) can be partially tested (fully tested if all $z_{\bar{T}} \in \Omega$ are observed) by comparing period t outcomes from treatment regimens identical through period t and different thereafter (although we do not know of any instance where this has been done). In balanced designs, where the same regimen is administered to all subjects, it is untestable (without additional assumptions).

Assumption (A3) implies that a model for the BOLD response at period t will depend only on the treatment history up to that period. Assumptions (A4) and (A5) constrain the manner in which the period t response depends on this history:

(A4) P Period Carry-Over — Let $0 \leq P \leq T - 1$ denote the smallest integer such that $z_{t-P} = z^*_{t-P}, \ldots, z_t = z^*_t$ implies $\Psi_{ivt}(z_{\bar{t}}) = \Psi_{ivt}(z^*_{\bar{t}})$ for all $t \geq P + 1$.

(A5) No Treatment by Period Interaction — For P in (A4) and $P + 1 \leq t < t'$, $z_{t-P} = z_{t'-P}, \ldots, z_t = z_{t'}$ implies $\Psi_{ivt'}(z_{t'}) = \Psi_{ivt}(z_{\bar{t}}) + c(t, t')$.

Assumption (A4) allows for P period carry over (with $P = 0$ corresponding to no carry-over) of the "true" potential outcomes $\Psi_{ivt}(z_{\bar{t}})$. The assumption is most useful when P is "small" compared to T. In conjunction with assumption (A2), $c(t, t')$ in assumption (A5) is 0; thus, any two regimens with the same treatments in the last P periods will have the same true scores; the BOLD responses, however, need not be the same.

In randomized cross-over studies, researchers often choose the time between successive treatments so that it is reasonable to assume there is no "carry-over" from prior treatments, formalizing this assumption by including only current treatment in models for period t responses. If one cannot make this assumption, limited amounts of carry-over, typically one period, are assumed.

To assure the validity of the no-carry over assumption in fMRI studies, an interval of approximately 30 seconds, the time it takes the BOLD contrast to return to baseline after a stimulus is applied, described by the hemodynamic response function (Lindquist, 2008) (see Figure 1A), would be required. But if intervals of such duration were used, it would severely limit the amount of data that could be collected from any given subject and/or substantially raise data collection costs. Consequently, inter-stimulus durations are typically much shorter. Assumption (A4) compensates for this, allowing period t outcomes to depend on current treatment and treatment in P previous periods. This necessitates choosing P. The relationship between the BOLD response and neural activity is known to be well approximated by a linear time invariant system with a hemodynamic response function known up to amplitude; this is used to choose P in empirical work and model the BOLD responses (as below).

In randomized cross-over studies, the no carry-over assumption is often combined with the assumption that treatment effects are invariant over periods. For the case of P period carry over, assumption (A5) (stated in terms of outcomes $\Psi_{ivt}(z_{\bar{t}})$) implies temporal invariance of

treatment effects: if

$$\{\mathbf{z}_{t-p}\}_{p=0}^{P}=\left\{\mathbf{z}_{t'-p}\right\}_{p=0}^{P},\left\{\mathbf{z}_{t-p}^{*}\right\}_{p=0}^{P}=\left\{\mathbf{z}_{t'-p}^{*}\right\}_{p=0}^{P},\psi_{ivt}\left(\bar{\mathbf{z}}_{t},\bar{\mathbf{z}}_{t}^{*}\right)=\psi_{ivt}\left(\bar{\mathbf{z}}_{t'},\bar{\mathbf{z}}_{t'}^{*}\right).$$

In general, with longitudinal data, assumptions (A4) and (A5) would be (fully) testable if all regimens of interest were observed; for the "small" *n*, "large" *R* case, including the case of the balanced design, (A4) and (A5) can be assessed by comparing alternative models for the response.

3.2 Models

fMRI researchers are typically interested in the average amplitude and variance in amplitude of a response to treatment. We model these quantities and relate them to the quantities in equations (1)-(4) using a linear mixed model for the potential outcomes in that implies assumptions (A1)-(A5):

$$Y_{iv,k_i+t}\left(\bar{\mathbf{z}}_{t}\right)=\boldsymbol{\alpha}_{v}+\mathbf{a}_{iv}+(\boldsymbol{\beta}_{v}+\mathbf{b}_{iv})'\mathbf{f}_{t}\left(\bar{\mathbf{z}}_{t}\right)+(\boldsymbol{\gamma}_{1v}+\mathbf{g}_{i1v})'\mathrm{N}_{t}+(\boldsymbol{\gamma}_{2v}+\mathbf{g}_{i2v})'\mathrm{N}_{i,k_i+t}\left(\bar{\mathbf{z}}_{t}\right)+\varepsilon_{iv,k_i+t}\left(\bar{\mathbf{z}}_{t}\right), \quad (5)$$

where $\alpha_v + a_{iv}$ is a random intercept, $f_t(\bar{z_t}) = (f_1(\bar{z_t}), ..., f_J(\bar{z_t}))'$, $\alpha_v + a_{iv} + (\beta_v + b_{iv})'f_t(\bar{z_t}) = \Psi_{ivt}(\bar{z_t})$ in (1), $f_j\left(\bar{\mathbf{z}}_{t}\right)=\sum_{p=0}^{P}z_{j,t-p}h_p$ is the convolution of the treatment subsequence $(z_{j,t-P}, ..., z_{jt})$ with the hemodynamic response function *h* (Lindquist, 2008), $(\gamma_{1v} + g_{i1v})'N_t + (\gamma_{2v} + g_{i2v})'N_{i,ki+t}(\bar{z_t})$ is the task dependent systematic error $B_{iv,ki+t}(\bar{z_T})$ in (1), $(N'_t, N'_{i,ki+t}(\bar{z_t}))'$ is a vector of measured "nuisance covariates",

$\mathsf{E}\left(\varepsilon_{iv,k_i+t}\left(\bar{\mathbf{z}}_{t}\right)|\boldsymbol{\theta}_{iv},\left\{\mathrm{N}_{i,k_i+t}\left(\bar{\mathbf{z}}_{t}\right)\right\}_{t=1}^{T}\right)=0$, where $\theta_{iv} = (a'_{iv}, b'_{iv}, g'_{i1v}, g'_{i2v})'$ and

$\mathsf{E}\left(\boldsymbol{\theta}_{iv}|\left\{\mathrm{N}_{i,k_i+t}\left(\bar{\mathbf{z}}_{t}\right)\right\}_{t=1}^{T}\right)=0$. Thus, the subject and average causal effects (2) and (3) are, respectively:

$$\psi_{ivt}\left(\bar{\mathbf{z}}_{t},\bar{\mathbf{z}}_{t}^{*}\right)=(\boldsymbol{\beta}_{v}+\mathbf{b}_{iv})'\left(\mathbf{f}_{t}\left(\bar{\mathbf{z}}_{t}\right)-\mathbf{f}_{t}\left(\bar{\mathbf{z}}_{t}^{*}\right)\right), \quad (6)$$

$$E\left(\psi_{ivt}\left(\bar{\mathbf{z}}_{t},\bar{\mathbf{z}}_{t}^{*}\right)\right)=\boldsymbol{\beta}_{v}'\left(\mathbf{f}_{t}\left(\bar{\mathbf{z}}_{t}\right)-\mathbf{f}_{t}\left(\bar{\mathbf{z}}_{t}^{*}\right)\right), \quad (7)$$

and the variance of the subject effects is

$$V\left(\psi_{ivt}\left(\bar{\mathbf{z}}_{t},\bar{\mathbf{z}}_{t}^{*}\right)\right)=\left(\mathbf{f}_{t}\left(\bar{\mathbf{z}}_{t}\right)-\mathbf{f}_{t}\left(\bar{\mathbf{z}}_{t}^{*}\right)\right)'\Sigma_{v\mathbf{bb}}\left(\mathbf{f}_{t}\left(\bar{\mathbf{z}}_{t}\right)-\mathbf{f}_{t}\left(\bar{\mathbf{z}}_{t}^{*}\right)\right), \quad (8)$$

where Σ_{vbb} is the covariance matrix of b_{iv}.

The effects (6) and (7) comparing different regimens are composed of the sums of the effects of individual stimuli in the *P* periods prior to and including *t*. In particular, the effect of stimulus *j* is measured by the amplitude β_{jv} ($\beta_{jv} + b_{ijv}$), the j^{th} component of β_v ($\beta_v + b_{iv}$);

this, the average effect (effect for subject *i*) of a 1 unit difference in the j^{th} component of $f_t(z_t) - f_t(z^*_t)$, is not affected by whether or not other stimuli are activated.

The sources of systematic error that fMRI researchers sometimes adjust for include 1) scanner drift and 2) head motion related artifacts not accounted for during preprocessing. Scanner drift is a low frequency change in the MR signal due to the imaging hardware, which creates slow changes in voxel intensities, and thus slow changes in the BOLD response. This source of systematic error does not depend on the treatment regimen $z_{\bar{T}}$ and depends on the subject only through $k_i + t$, $t = 1, ..., T$. Drift is often modeled as

$$D_{iv,k_i+t} = (\boldsymbol{\gamma}_{1v} + \mathbf{g}_{i1v})' \mathrm{N}_t + \delta_{iv,k_i+t}. \quad (9)$$

where $E\left(\delta_{iv,k_i+t} | \boldsymbol{\theta}_{iv}, \left\{\mathrm{N}_{i,k_i+t}\left(\bar{\mathbf{z}}_t\right)\right\}_{t=1}^{T}\right) = 0$ and the ℓ^{th} coordinate of the basis vector N_t, $\ell = 1, \ldots, L_1$, is, typically, $N_{\ell t} = t^{\ell}$. (The term t^0 cannot be included in the model for drift, as (5) already includes random intercepts; note that incorporating drift constants into the subject intercepts does not affect estimates of the treatment effects and allows for greater comparability with drift models that use other basis vectors, e.g., discrete cosine basis sets.)

It is generally held that the rigid body motion-correction procedures typically used in the pre-processing of fMRI data are effective at aligning voxels to their appropriate position in three dimensional space. But these procedures do not account for motion related changes in the physical properties of the nuclei being scanned, which give rise to a 'spin-history' artifact. To account for this, some researchers include motion regressors in models for the BOLD response:

$$H_{iv,k_i+t}\left(\bar{\mathbf{z}}_t\right) = (\boldsymbol{\gamma}_{2v} + \mathbf{g}_{i2v})' \mathrm{N}_{i,k_i+t}\left(\bar{\mathbf{z}}_t\right) + \kappa_{iv,k_i+t}\left(\bar{\mathbf{z}}_t\right), \quad (10)$$

where $E\left(\kappa_{iv,k_i+t}\left(\bar{\mathbf{z}}_t\right) | \boldsymbol{\theta}_{iv}, \left\{\mathrm{N}_{i,k_i+t}\left(\bar{\mathbf{z}}_t\right)\right\}_{t=1}^{T}\right) = 0$ and $\mathrm{N}_{i,ki+t}(z_{\bar{t}})$ is a vector of "nuisance covariates" for head motion for subject *i* at voxel *v* at time $k_i + t$. Typically the motion regressors are six time courses corresponding to three translations and three rotations of the brain computed using a rigid-body transformation for each functional scan (see Figure 1C for example); thus, the values do not depend on *v*.

Unlike scanner drift, systematic error due to head motion may depend on the treatment regimen $z_{\bar{T}}$, and while a minority of researchers include motion regressors to correct for head movement, no principled justification for this practice has been put forth in the neuroimaging literature. In randomized experiments with $T = 1$, it is well known that adjustment for intermediate outcomes affected by treatment leads to biased estimates of causal effects (Rosenbaum, 1984). Here, however, if motion regressors are not included and they are correlated with $f_t(z_{\bar{t}})$, estimates of the amplitudes β_v will be biased, hence estimates of the unit and average effects (2) and (3) will also be biased.

Additional motion due to other sources $\eta_{iv,ki+t}(\bar{z}_t)$, e.g., cardiac and respiratory activity, is typically assumed to be random, with mean 0, therefore included in the error term $\varepsilon_{iv,ki+t}(\bar{z}_t)$; thus

$$\varepsilon_{iv,k_i+t}\left(\bar{\mathbf{z}}_t\right)=\eta_{iv,k_i+t}\left(\bar{\mathbf{z}}_t\right)+\delta_{iv,k_i+t}+\kappa_{iv,k_i+t}\left(\bar{\mathbf{z}}_t\right). \quad (11)$$

To complete the model, we specify the distribution of the random components and the relationship among these. Let $\varepsilon_{iv}(\bar{z}_T) = (\varepsilon_{iv,ki+1}(\bar{z}_1), ..., \varepsilon_{iv,ki+T}(\bar{z}_T))'$. We assume $\varepsilon_{iv}\left(\bar{\mathbf{z}}_T\right)|\boldsymbol{\theta}_{iv},\left\{\mathrm{N}_{i,k_i+t}\left(\bar{\mathbf{z}}_t\right)\right\}_{t=1}^{T}\sim\mathcal{N}\left(\mathbf{0},\boldsymbol{\Sigma}_{\mathbf{v},\varepsilon_i\varepsilon_i}\left(\bar{\mathbf{z}}_\mathbf{T}\right)\right)$, where $\mathcal{N}$ denotes the normal distribution, $\left(\boldsymbol{\theta}_{iv}|\left\{\mathrm{N}_{i,k_i+t}\left(\bar{\mathbf{z}}_t\right)\right\}_{t=1}^{T}\right)\sim\mathcal{N}\left(\mathbf{0},\boldsymbol{\Sigma}_{\mathbf{v}\boldsymbol{\theta}\boldsymbol{\theta}}\right)$, with $\mathrm{E}\left(\mathbf{g}_{1iv}\mathbf{b}_{iv}'|\left\{\mathrm{N}_{i,k_i+t}\left(\bar{\mathbf{z}}_t\right)\right\}_{t=1}^{T}\right)=0$, $E\left(\mathbf{g}_{1iv}\mathbf{g}_{2iv}'|\left\{\mathrm{N}_{i,k_i+t}\left(\bar{\mathbf{z}}_t\right)\right\}_{t=1}^{T}\right)=0$. The random vectors $\left(\boldsymbol{\varepsilon}_{iv}',\boldsymbol{\theta}_{iv}'\right)'|\left\{\mathrm{N}_{i,k_i+t}\left(\bar{\mathbf{z}}_t\right)\right\}_{t=1}^{T}$, $i = 1, ..., n$, are assumed independent, with θ_{iv} and $\varepsilon_{iv}(\bar{z}_T)$ independent, given $\left\{\mathrm{N}_{i,k_i+t}\left(\bar{\mathbf{z}}_t\right)\right\}_{t=1}^{T}$.

In functional neuroimaging experiments, subjects are typically randomly assigned to regimens in Ω or, as in the case of the balanced design, assigned with probability one to a particular regimen. In both cases, it is easy to see $E(Y_{iv,ki+t}(\bar{z}_t)) = E(Y_{iv,ki+t} \mid \bar{z}_t = \bar{z}_t)$; thus estimation of the causal model (5) is straightforward, using for example, maximum likelihood (ML). Subject effects (2) are then estimated as: $\hat{\psi}_{ivt}\left(\bar{\mathbf{z}}_t,\bar{\mathbf{z}}_t^*\right)=\left(\hat{\boldsymbol{\beta}}_v+\hat{\mathbf{b}}_{iv}\right)'\left(\mathbf{f}_t\left(\bar{\mathbf{z}}_t\right)-\mathbf{f}_t\left(\bar{\mathbf{z}}_t^*\right)\right)$, where $\hat{\boldsymbol{\beta}}_v$ is the estimate of β_v and $\hat{b_{iv}}$ the predictor of b_{iv}. Average causal effects (3) are estimated replacing β_v in (7) with the estimate $\hat{\boldsymbol{\beta}}_v$ and the variance in the subject effects is estimated using the estimate $\hat{\Sigma}_{v\mathbf{bb}}$ in place of $\Sigma_{v\mathrm{bb}}$ in (8).

An important caveat is in order. While limited types of departures from the causal model (5) can be assessed by adding additional terms, the model parameters are invariant over regimens. When subjects are randomly assigned to different regimens, it is possible to examine this assumption by comparing subjects assigned to different regimens, but in a balanced design, this is not possible.

4 Application to SET Task Data Set

Twenty-five healthy, right-handed, native English speaking undergraduates were recruited at the University of Michigan as subjects and scanned with BOLD fMRI at 3T (GE, Milwaukee, WI). A series of 215 images was acquired using a T2*-weighted, single-shot reverse spiral acquisition (gradient echo, TR= 2000, TE= 30, flip angle= 90°) with 40 sequential axial slices (FOV= 20, 3.12 × 3.12 × 3 mm, 64 × 64 matrix). High-resolution T1 spoiled gradient recall (SPGR) images were acquired for anatomical localization and

warping to standard space. Functional images were subjected to a standard preprocessing sequence. Slice-timing acquisition correction was performed using sync interpolation and realignment of the functional images to correct for head movement was performed using the Automated Image Registration tools (Woods et al., 1998). The remaining preprocessing steps were performed using the Statistical Parametric Mapping analysis package (SPM2, Wellcome Department of Cognitive Neurology, London, UK). Spoiled Gradient Echo (SPGR) images (i.e., structural scans) for each participant were co-registered to the mean functional image. The SPGR images were then normalized to the anatomical space of the 152-brain template provided by the Montreal Neurological institute (MNI), providing a mapping between the functional images and the brain template. The mapping parameters were then applied to each functional image to align the image with a standard stereotaxic template space, thereby allowing inter-subject comparisons. Finally, the normalized functional images were smoothed with an 8-mm full-width at half-maximum Gaussian smoothing kernel.

To maintain consistency with the way neuroscientists analyze fMRI data, we employ the "massively univariate approach", estimating the model of equation (5) at each voxel. However, we do not use the ad-hoc approach, common in the neuroimaging literature, of estimating the model for each subject and then combining the results across subjects using a simple average of the subject results. Instead, we estimate the mixed-effects model constructed in the previous section.

As previously noted, there has been some debate in the literature as to whether or not to include motion regressors in models of the BOLD response, with some arguing that inclusion is necessary to reduce signal bias due to data artifacts, others arguing that inclusion may induce bias by reducing signal magnitude. In a randomized experiment with $T = 1$, it is well known that adjusting for post treatment intermediate outcomes affected by the treatment leads to biased estimates of treatment effects (Rosenbaum, 1984), suggesting that stimulus dependent motion regressors should not be included in models for the BOLD response. However, when the outcome is measured with stimulus dependent systematic error, as here, treatment effects defined using the BOLD responses $Y_{iv,ki+t}(\bar{z_T})$ will evidence both causation and error; thus we define these effects using the "true" responses $\Psi_{ivt}(\bar{z_T})$ and accordingly adjust for the task dependent motion regressors. Our resolution of this issue is important, as (see below) substantive conclusions can depend on whether or not these regressors are included.

To illustrate the impact of adjustment in the SET task study, we fit two models. Both include (a) two activation regressors: 1) "EarlyPrep" for activation from the first visual cue to the middle of speech preparation ($z_{1t} = 1$ for $t \in [61, 90]$, 0 otherwise) and 2) "LatePrep" for activation from the middle of preparation until the presentation of the relief cue ($z_{2t} = 1$ for t *2* [91, 120], 0 otherwise), and the drift regressors t and t^2. Similar activation regressors are used by Waugh et al. (2012) in the analysis of a related stress task for a sample of subjects suffering from depression disorder. Model 1 also includes the 6 motion regressors measuring how far the brain lies from a reference image at each time point.

Figure 1 illustrates the model components. Figure 1A depicts the canonical hemodynamic response function *h* widely used (and used here) in modeling the BOLD response. *h* is the change in the response (up to amplitude) of the fMRI signal associated with the neuronal activity triggered by a short stimulus; the (approximately) 30 seconds it takes for the signal to return to baseline is the carry-over, leading to our choice of $P = 15$ two second intervals. Figure 1B depicts the time course of the "EarlyPrep" and "LatePrep" variables. Figure 1C displays the head movement covariates from realignment for a single subject at each of the 215 functional images. Since we used a rigid body transformation, the 6 terms represent displacement in the *x*, *y* and *z* directions and rotational displacement ('pitch', 'roll' and 'yaw') in 3 dimensional space for each functional image.

Maximum likelihood was used to estimate the model of equation (5) (both with and without motion regressors). The heteroscedastic errors were assumed to follow an AR(1) structure. Variance components were estimated using restricted maximum likelihood. The models were fit using custom Matlab code that implemented the restricted iterative generalized least squares (RIGLS) algorithm (Goldstein (1989)). Lindquist et al. (2012) showed that estimates obtained using RIGLS are comparable in performance and computational demands to those using linear mixed effects (LMER) in R. Estimates of fixed effects were obtained at each voxel, along with the variance of the subject effects. As is common in the literature, a cluster forming threshold (Poldrack et al., 2011) was used to locate areas of activation. First, for each of the activation regressors, a t-test was performed at each voxel, using a significance level (primary threshold) of = .01. Next, a secondary threshold of size *k* ($k = 10$ here) is chosen, with contiguous groups of *k* or more voxels signficant at $\alpha = .01$ deemed active. We also examined several other values for and *k* ($\alpha \in (0.001, 0.05)$, $k \in (0, 20)$), and the substantive results did not change.

Thresholded t-maps for the parameter associated with "EarlyPrep" obtained using Model 1 are presented for a single slice (number 18 out of 46) in the left panel of Figure 2. Voxels are color-coded according to the value of the associated t-statistic; orange represents barely significant results, while white represents the maximum value ($t \approx 10$). There is clear activation in the visual cortex during the "EarlyPrep" period, expected as this period coincides with the presentation of the visual cue. The superior temporal cortices associated in social neuroscience studies with inferences about agency, among other things, were also activated during the first part of the task. Finally, the ventromedial prefrontal cortex, an area associated with visceromotor control, self-related attention, and generation and regulation of emotion based on context showed sustained activation throughout the speech preparation. The results are consistent with those in Robinson et al. (2010), who performed an analysis using regions of interest, rather than the voxel-wise analysis here.

The right panel of Figure 2 displays results for Model 2, illustrating the bias that occurs when motion regressors are excluded. Under Model 2, less active voxels are found in both the visual and ventromedial prefrontal cortex, and no significant activation is found in the superior temporal cortices. This bias is further illustrated in Figure 3, where the difference in estimates of the "Early Prep" coefficient between Models 1 and 2 for five equally spaced slices (8, 16, 24, 32 and 40 out of a total of 46 slices) show significantly higher values for Model 1 in key regions associated with stress. More generally, for other slices (not shown

here) there is consistent activation in the ventricles and at the edge of the brain, symptomatic of motion-related artifacts.

Individual differences in effects are reflected in the variance (4). Figure 4 displays estimated variances for the “Early Prep” random effects under Model 1. The between subject variability seems small in the visual cortex, indicating that subjects process simple visual instructions in a similar manner. In contrast, the variability is greater in the ventromedial prefrontal cortex, suggesting larger individual differences in the way emotion is regulated in response to the stress task.

5 Discussion

Treatment effects are defined as comparisons of subjects’ measured outcomes (only one of which is observed) under different treatment regimens. But if these outcomes are measured with stimulus dependent systematic error, as in fMRI experiments, such definitions will conflate causation with measurement error. Therefore we differentiate these components of measured outcomes and define effects free of systematic error. Given this definition, to consistently estimate these effects it is necessary to adjust for task dependent systematic error, even though it is affected by the treatment.

In the literature on longitudinal causal inference, average treatment effects are identified using ignorability conditions that postulate the independence of potential outcomes and treatments (possibly conditional on covariates, past outcomes and treatments) in conjunction with positivity conditions ensuring positive probability of observing responses under different treatment regimens. Often the number of possible regimens far exceeds the number of subjects, precluding estimation by direct comparison of similar types of subjects exposed to different regimens. Therefore, estimation is model based. Under the assumption that model parameters are the same in all regimens of interest, extrapolation from observed to unobserved regimens is warranted. This assumption can be addressed (in part) when multiple regimens are observed, but in the case of a balanced design, as herein, it is not possible to do so.

We analyze data from a study of social evaluative threat. A linear mixed model is used to estimate unit effects, average effects, and the variance of the unit effects. Excluding motion covariates leads to biased estimates, with spurious activations around the edges of the brain and the ventricles. Under Model 1, which includes the motion covariates, we find that the visual cortex and superior temporal cortices are activated during the early part of speech preparation and the ventromedial prefrontal cortex is activated in both early and later phases of the task. In addition, we find little between subject variation in activation in the visual cortex, greater variation in the ventromedial prefrontal cortex.

Our approach and results have implications more generally for the manner in which neuroscientists make causal inferences using fMRI data. Currently, less than a third of analyses in the literature include motion regressors in models for the BOLD respones (Carp, 2012). Our framework (supported by the empirical results) indicates that motion regressors

should be included in models for the BOLD response to adjust for task dependent systematic error.

In future work, we will extend our framework in a number of directions. As an example, whereas we have modeled the data at each voxel and inferred clusters of activated voxels, in lieu of this *a posteriori* approach, neuroscientists sometimes define brain regions using clusters of voxels chosen *a priori*, and perform a region of interest (ROI) analysis, in which the goal is to characterize the relationship between a stimulus and activity in the region. These analyses are typically performed by averaging a subject's BOLD responses over the voxels comprising the region and modelling the aggregated responses. However, unless the systematic errors average out to 0 over the region, and there is no reason to think that will be so, the averaged responses will also contain systematic error. In general, such an analysis will both lead to biased estimates of treatment effects and also obscure any variation in signal across the region. Such variation, if present, may be of interest in localizing subregions that account for the allocation and processing of resources within the ROI. A better approach, conceptually straightforward, but computationally demanding, is to extend the model of Section 3 over the voxels in the predefined region, treating the amplitudes as random effects over voxels and subjects: if the signal is homogeneous throughout the region, the null hypothesis that the voxel variance is 0 will not be rejected (Lindquist et al., 2012). As a second example, Lindquist (2012) modeled the relationship between a thermal stimulus and reported pain, as mediated by the time course of activity, treated as a functional mediator, in various brain regions. The BOLD responses $Y_{iv,ki+t}(z_t)$ in the region were averaged and the averages treated as the components of the functional mediator. As above, the averaged responses will generally reflect both task related effects and task dependent systematic error, leading to biased estimates of treatment effects. In future work, we shall separate these components, defining the components of the functional mediator using the "true" hemodynamic responses $\Psi_{ivt}(z_t)$.

Acknowledgments

We thank the Editor and two anonymous reviewers for remarks that helped us improve this paper, and Tor Wager and Christian Waugh for supplying the data. This research was supported by NIH grant R01EB016061.

References

Bassett DS, Bullmore E. Small-world brain networks. The Neuroscientist. 2006; 12(6):512–523. [PubMed: 17079517]

Carp J. The secret lives of experiments: Methods reporting in the fmri literature. NeuroImage. 2012; 63:289–300. [PubMed: 22796459]

Churchill NW, Oder A, Abdi H, Tam F, Lee W, Thomas C, Ween JE, Graham SJ, Strother SC. Optimizing preprocessing and analysis pipelines for single-subject fmri. i. standard temporal motion and physiological noise correction methods. Human Brain Mapping. 2012; 33(3):609–627. [PubMed: 21455942]

Friston K, Harrison L, Penny W. Dynamic causal modelling. NeuroImage. 2003; 19:1273–1302. [PubMed: 12948688]

Friston KJ, Holmes AP, Poline J, Grasby P, Williams S, Frackowiak RS, Turner R. Analysis of fmri time-series revisited. NeuroImage. 1995; 2(1):45–53. [PubMed: 9343589]

Goldstein H. Restricted unbiased iterative generalized least-squares estimation. Biometrika. 1989; 76(3):622–623.

Johnstone T, Ores Walsh KS, Greischar LL, Alexander AL, Fox AS, Davidson RJ, Oakes TR. Motion correction and the use of motion covariates in multiple-subject fmri analysis. Human Brain Mapping. 2006; 27(10):779–788. [PubMed: 16456818]

Lindquist M. Functional causal mediation analysis with an application to brain connectivity. Journal of the American Statistical Association. 2012; 107:1297–1309. [PubMed: 25076802]

Lindquist M, Sobel M. Graphical models, potential outcomes and causal inference: Comment on Ramsey, Spirtes and Glymour. NeuroImage. 2011; 57:334–336. [PubMed: 20970507]

Lindquist M, Sobel M. Cloak and DAG: A response to the comments on our comment. NeuroImage. 2013; 76:446–449. [PubMed: 22119004]

Lindquist MA. The statistical analysis of fMRI data. Statistical Science. 2008; 23:439–464.

Lindquist MA, Spicer J, Asllani I, Wager TD. Estimating and testing variance components in a multi-level glm. NeuroImage. 2012; 59(1):490–501. [PubMed: 21835242]

Lord, FM.; Novick, MR. Statistical theories of mental test scores. 1968.

Lund TE, Madsen KH, Sidaros K, Luo W-L, Nichols TE. Non-white noise in fmri: does modelling have an impact? NeuroImage. 2006; 29(1):54–66. [PubMed: 16099175]

Luo X, Small D, Li C, Rosenbaum P. Inference with interference between units in an fMRI experiment of motor inhibition. Journal of the American Statistical Association. 2012; 107:530–541.

McIntosh A, Gonzalez-Lima F. Structural equation modeling and its application to network analysis in functional brain imaging. Human Brain Mapping. 1994; 2:2–22.

Poldrack, R.; Mumford, J.; Nichols, T. Handbook of Functional MRI Data Analysis. Cambridge University Press; 2011.

Ramsey J, Hanson S, Hanson C, Halchenko Y, Poldrack R, Glymour C. Six problems for causal inference from fMRI. NeuroImage. 2010; 49:1545–1558. [PubMed: 19747552]

Robins, J.; Hernán, M. Estimation of the effects of time-varying exposures.. In: Fitzmaurice, G.; Davidian, M.; Verbeke, G.; Molenberghs, G., editors. Longitudinal Data Analysis. Chapman and Hall/CRC; 2009. p. 553-597.

Robins JM, Greenland S, Hu F-C. Estimation of the causal effect of a time-varying exposure on the marginal mean of a repeated binary outcome. Journal of the American Statistical Association. 1999; 94(447):687–700.

Robinson LF, Wager TD, Lindquist MA. Change point estimation in multi-subject fmri studies. NeuroImage. 2010; 49(2):1581–1592. [PubMed: 19733671]

Roebroeck A, Formisano E, Goebel R. Mapping directed influence over the brain using Granger causality and fMRI. NeuroImage. 2005; 25:230–242. [PubMed: 15734358]

Rosenbaum P. The consquences of adjustment for a concomitant variable that has been affected by the treatment. Journal of the Royal Statistical Society. Series A (General). 1984; 147:656–666.

Rosenbaum PR, Rubin DB. The central role of the propensity score in observational studies for causal effects. Biometrika. 1983; 70(1):41–55.

Rubin D. Comment on ”Randomization analysis of experimental data: The Fisher randomization test”, by D. Basu. Journal of the American Statistical Association. 1980; 75:591–593.

Tracey I. Imaging pain. British Journal of Anaesthesia. 2008; 101(1):32–39. [PubMed: 18556697]

Valdes-Sosa P, Roebroeck A, Daunizeau J, Friston K. Effective connectivity: Influence, causality and biophysical modeling. NeuroImage. 2011; 58:339–361. [PubMed: 21477655]

Wager T, Waugh C, Lindquist M, Noll D, Fredrickson B, Taylor S. Brain mediators of cardiovascular responses to social threat, Part I: Reciprocal dorsal and ventral sub-regions of the medial prefrontal cortex and heart-rate reactivity. NeuroImage. 2009; 47:821–835. [PubMed: 19465137]

Waugh CE, Hamilton JP, Chen MC, Joormann J, Gotlib IH, et al. Neural temporal dynamics of stress in comorbid major depressive disorder and social anxiety disorder. Biology of mood & anxiety disorders. 2012; 2(1):11. [PubMed: 22738335]

Woods RP, Grafton ST, Holmes CJ, Cherry SR, Mazziotta JC. Automated image registration: I. general methods and intrasubject, intramodality validation. Journal of Computer Assisted Tomography. 1998; 22(1):139–152. [PubMed: 9448779]

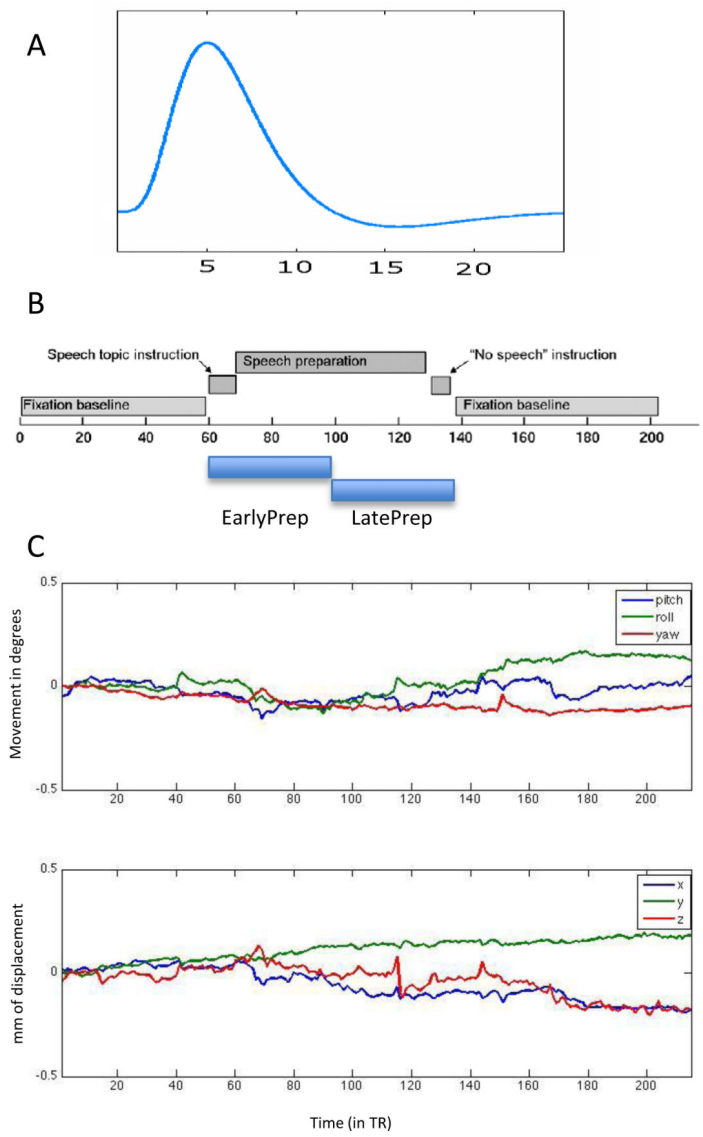

Figure 1.

(A) The hemodynamic response function describes the time course of the BOLD signal after a brief stimulation. (B) The experimental design, shown together with the timing of the "EarlyPrep" and "LatePrep" treatments. (C) Plots of the six motion covariates obtained from performing rigid body motion correction on a single subject for each of the 215 functional images. The top panel shows rotations in three-dimensional space across time. The bottom panel shows translation in the *x*, *y* and *z* directions across time.

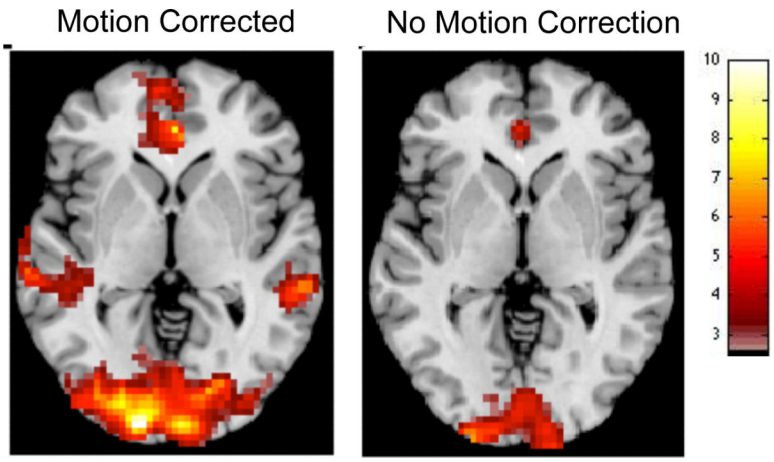

Figure 2.
Thresholded t-maps for the parameter associated with “EarlyPrep” obtained using Model 1 (left) and Model 2 (right).

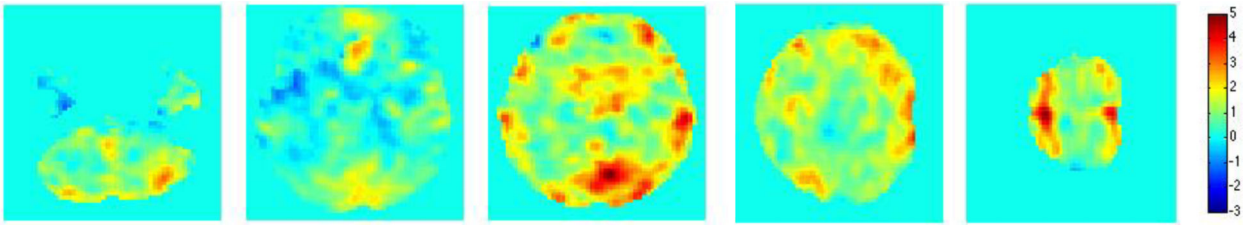

Figure 3.
Differences between “EarlyPrep” coefficients in Models 1 and 2 for five equally spaced slices (slice numbers 8, 16, 24, 32 and 40).

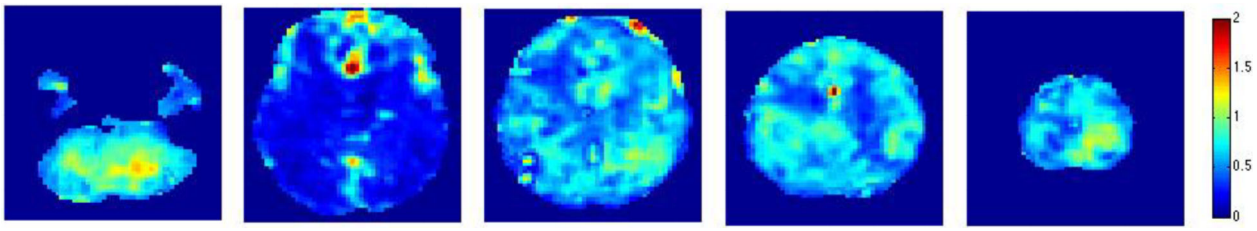

Figure 4.
Estimated between subject variance for “EarlyPrep” random effects under Model 1 for five equally spaced slices (slice numbers 8, 16, 24, 32 and 40).