The Annals of Applied Statistics
2020, Vol. 14, No. 1, 452–472
https://doi.org/10.1214/19-AOAS1316

ESTIMATING CAUSAL EFFECTS IN STUDIES OF HUMAN BRAIN FUNCTION: NEW MODELS, METHODS AND ESTIMANDS

BY MICHAEL E. SOBEL[1] AND MARTIN A. LINDQUIST[2]

[1]*Department of Statistics, Columbia University, michael@stat.columbia.edu*
[2]*Department of Biostatistics, Johns Hopkins University, mlindqui@jhsph.edu*

Neuroscientists often use functional magnetic resonance imaging (fMRI) to infer effects of treatments on neural activity in brain regions. In a typical fMRI experiment, each subject is observed at several hundred time points. At each point, the blood oxygenation level dependent (BOLD) response is measured at 100,000 or more locations (voxels). Typically, these responses are modeled treating each voxel separately, and no rationale for interpreting associations as effects is given. Building on Sobel and Lindquist (*J. Amer. Statist. Assoc.* **109** (2014) 967–976), who used potential outcomes to define unit and average effects at each voxel and time point, we define and estimate both "point" and "cumulated" effects for brain regions. Second, we construct a multisubject, multivoxel, multirun whole brain causal model with explicit parameters for regions. We justify estimation using BOLD responses averaged over voxels within regions, making feasible estimation for all regions simultaneously, thereby also facilitating inferences about association between effects in different regions. We apply the model to a study of pain, finding effects in standard pain regions. We also observe more cerebellar activity than observed in previous studies using prevailing methods.

1. Introduction. Functional magnetic resonance imaging (fMRI) (Kwong et al. (1992), Ogawa et al. (1990)) is a noninvasive procedure for whole brain imaging with good spatial resolution and in which neuronal activity is measured indirectly through changes in brain hemodynamics. In a typical fMRI study, each subject is administered one or more stimuli and observed at hundreds of time points. At each time point, the subject's blood oxygenation level dependent (BOLD) response is recorded at roughly 100,000 spatial locations (voxels), yielding multivariate time series data (Li (2014)). Localization studies estimate the effects of one or more stimuli on brain activity in different locations. Here, the association between experimental stimuli and BOLD responses is typically modeled voxel by voxel (Lindquist (2008)) or the BOLD responses are averaged over predefined regions of interest (ROIs), defined as a collection of adjacent voxels, then modeled separately by region. (Poldrack (2007)). Parameter estimates describing the association are then deemed effects.

Increasingly, researchers are also interested in effective connectivity, that is, the integration of neural activity among different brain regions and the causal relations among activity in these areas (Friston (2011), Lindquist and Sobel (2016)). To study this, neuroscientists estimate the association between BOLD responses in different voxels (or averages of BOLD responses within regions) using various statistical methods, for example, Granger causal mapping (Roebroeck, Formisano and Goebel (2005)), dynamic causal modeling (Friston, Harrison and Penny (2003)), structural equations and directed graphical models (Mclntosh and Gonzalez-Lima (1994)) and, following common practice, interpret parameter estimates as effects.

Received July 2019; revised November 2019.
Key words and phrases. Causal inference, fMRI, functional connectivity, systematic error, pain, region of interest.

The approaches to causal inference above lack foundation. Researchers typically do not indicate what they mean by causation nor the manner in which and/or conditions under which estimated associations support a causal interpretation. To provide foundation, Lindquist and Sobel (2011, 2013) advocated using the potential outcomes framework from the statistical literature on causal inference in fMRI research.

Sobel and Lindquist (2014) (hereafter SL) start at the most elemental level, defining the unit effect of treatment sequence s vs. sequence s' for subject i at a specific voxel $v^{(b)}$ in brain region b at measurement time t. The unit effects cannot be observed directly, but these and their average (and variance) over a population of subjects are identified from the observed data under a model for the BOLD responses. Following the predominant "general linear model" (GLM) approach (Friston et al. (1994)), in which a separate model is estimated at each voxel, SL modeled the BOLD response at time t as the sum of a hemodynamic response function (HRF) describing the time course of blood flow to the brain, a systematic error and a random error. As is common, they modeled the signal, that is, the HRF, as the product of an unknown subject, voxel and treatment specific amplitude with a canonical HRF (CHRF), assumed to be known and invariant over subjects, voxels and treatments. As neural activity tends to cluster in ROIs composed of multiple voxels, and, as it is the activity in these ROIs that is of primary interest, SL, following common practice, used a cluster based thresholding procedure to group adjacent voxels into clusters (Poldrack, Mumford and Nichols (2011)) and heat maps color coded to correspond with the value of the associated t-statistic to display the effects.

The foregoing approach is problematic. SL define causal effects at the voxel level, then use the thresholding procedure to declare affected regions without ever defining causal effects for regions. The results are also sensitive to the thresholds chosen (Carp (2012), Woo, Krishnan and Wager (2014)). Further, while the assumption that the HRF is the product of an unknown amplitude with the known CHRF allows for direct comparisons of amplitudes across voxels, subjects and treatments, it is not biologically plausible (Monti (2011)), and its use will lead to biased estimates of activation and connectivity, as these depend on the model for the HRF. In addition, as heat maps display t ratios, rather than the amount of neural activity in a voxel or, by extension, within an ROI, neural activity in region A can exceed that in region B even if the associated heat map suggests otherwise. Nor are heat maps useful for understanding connectivity, as they do not indicate the strength of association between neural activity in different voxels or regions.

In lieu of ad hoc cluster thresholding, some researchers have constructed models to account for the spatial association between different voxels. Woolrich et al. (2004) and Penny, Trujillo-Barreto and Friston (2005) proposed single-subject Bayesian models that account for spatial relations among "nearby" voxels. Harrison and Green (2010) generalized the latter model. Bowman (2007) proposed a multisubject, multivoxel, linear mixed model for the BOLD responses in a single brain region using a "functional" distance metric to account for correlation among distant voxels.

Several multisubject multivoxel models build explicitly on the GLM approach. Bowman (2005) modeled voxelwise activations in empirically defined clusters of voxels, then used the estimated activations to model relationships among activations in a cluster with a spatial autoregressive model. Bowman et al. (2008) proposed a whole brain Bayesian hierarchical model, also using a two-stage estimation procedure. Sanyal and Ferreira (2012) and Mejia et al. (2017) also construct multisubject multivoxel Bayesian hierarchical models using a two-stage approach. Zhang et al. (2016) recently proposed a one step multisubject, multivoxel, nonparametric Bayesian model. They point out that, even if variational Bayes is used for inference, the huge amount of data generated by fMRI experiments may necessitate some form of data reduction, for example, summarizing the responses over a region before applying the model to the whole brain.

This paper aims to develop a principled framework for causal inference at higher levels of brain organization. First, using voxelwise unit effects as building blocks, we define unit causal effects for ROIs, using these to define causal estimands for ROI time courses; we also consider the variance of the effects and the association between these in different regions. Second, to estimate these effects, we construct a multisubject, multiregion, multirun, hierarchical whole brain model. As in some prior work, we analyze BOLD responses averaged over regions; unlike such work, in which this approach is justified out of computational necessity, mathematical justification is given (Appendix B). While this approach has been criticized for ignoring the spatial structure of relationships among different voxels and the possibly different levels of activation within an ROI (Bowman (2007)), the spatiotemporal structure of neural activity depends on the anatomical structure of the brain, short and long distance connections among neurons and the treatment under study, and we do not believe current knowledge permits specification of reasonable spatiotemporal models at a fine grained level; thus, an advantage of our estimation procedure is that it does not require specifying relationships among voxels within an ROI. Third, our procedure is computationally feasible for hundreds of subjects and regions, allowing us to work with finely delineated ROIs, thereby mitigating the criticism that potentially highly heterogeneous activity within ROIs is ignored. Fourth, we estimate the functional connectivity between effects for all pairs of ROIs, something that estimation procedures that can only handle one or a few ROIs cannot do. We also display our results graphically, using whole-brain maps of causal effects, and spatiotemporal correlation plots facilitating the investigation of temporally lagged relationships between ROIs.

Our approach bears some resemblance to that of Bowman et al. (2008). But the differences are substantial. In stage one, Bowman et al. (2008) used the GLM approach to estimate single subject activations at every voxel; in stage two, using a Bayesian hierarchical model, the estimated activations are decomposed as the sum of a fixed effect for ROIs with a mean *zero* random subject effect for region and a mean *zero* random effect (common to all subjects) for voxels in a region. The activations within ROIs are assumed equicorrelated. In our stage one, single subject region level activations are estimated without imposing a correlation structure for the voxel level activations within ROIs; in stage two, the activations are modeled as the sum of a fixed effect and a random subject effect. Second, Bowman et al. (2008) model the HRF as the product of a scalar amplitude with the canonical HRF (CHRF). To avoid the biased estimates that result from this approach, we use basis sets to model the HRF. Although many com**a**monly used sets do not adequately handle the complexities of the HRF (Lindquist and Wager (2007), Lindquist et al. (2009)), Degras and Lindquist (2014) demonstrated that cardinal B splines with a high order and sufficient number of knots recover the HRF well; here, we model the HRF with 15 cardinal B-splines of order 6. In Appendix C, we show this accurately recovers the HRF for a variety of HRFs featuring various durations and onsets, whereas the standard approach fails to do so for HRFs that are not "close" to the CHRF.

We illustrate our approach using a study of thermal pain, where noxious heat stimuli were applied at different temperatures to the left forearm of each of 33 subjects. For every ROI, we estimate an "integrated average effect." In addition to the effects in standard pain regions, we observe more cerebellar and visual activity than is usually observed in pain studies of this type. Our approach also allows us to estimate the lagged correlation between HRFs across the brain. We illustrate these relationships using a spatiotemporal correlation plot that pinpoints enhanced correlation between pain-related regions both in reaction to the thermal stimuli as well as in the time preceding pain reporting, signaling a potential correlation of activity across brain regions during "pain recall."

We proceed as follows. The experiment and data are described in Section 2. In Section 3 notation is introduced and causal effects for voxels and brain regions are defined. In Section 4

we set out the whole brain causal model and the methods used to estimate causal effects and make inferences about these. The thermal pain data are analyzed in Section 5. Section 6 concludes.

2. An fMRI study of thermal pain. 33 healthy, right-handed subjects completed the study (age 27.9 ± 9.0 years, 22 females); all gave informed consent. The Columbia University Institutional Review Board approved the study. For each subject, seven runs were administered during a single session. Each run consisted of between 58–75 trials. In each trial, thermal stimulations were delivered to the volar surface of the left inner forearm. Each stimulus lasted 12.5 seconds, with *three* second ramp-up, *two* second ramp-down periods and 7.5 seconds at the target temperature. Six temperatures, ranging from 44.3 to 49.3°C in increments of 1°C, were administered to each participant; for the analysis, these were grouped into warm (<46°) and hot (>46°) stimuli (Wager et al. (2013)). Each stimulus was followed by a 4.5 to 8.5 second prerating period, after which subjects rated their intensity of pain on a scale of *zero* to 100; in this paper, as interest centers on the hemodynamic responses to the thermal stimuli, we do not analyze these rating data. Each trial ended with a *five* to *nine* second resting period, followed by a new trial, or, if the trial terminated a run, a brief (one or more minutes) resting period followed by a new run.

For each subject, 1845 images were acquired using a 3T Philips Achieva TX scanner at Columbia University. Structural images were acquired using high-resolution T1 spoiled gradient recall (SPGR) images. Functional echo planar images (EPIs) were acquired with repetition time (TR) = 2000 ms, echo time (TE) = 20 ms, field of view = 224 mm, 64×64 matrix, $3 \times 3 \times 3$ mm^3 voxels, 42 interleaved slices, parallel imaging and sensitivity encoding (SENSE) factor 1.5. For each subject, structural images were coregistered to the mean functional image using the iterative mutual information-based algorithm in SPM8;[1] the images were then normalized to Montreal Neurological Institute (MNI) space using SPM8's generative segment-and-normalize algorithm. Prior to preprocessing of functional images, the first four volumes were removed to allow for image intensity stabilization. Outliers were identified using the Mahalanobis distance for the matrix of slicewise mean and standard deviation values. The functional images were corrected for differences in slice-timing, and the motion was corrected using SPM8. These images were warped to SPM's normative atlas using warping parameters estimated from coregistered high-resolution structural images, and smoothed with an 8 mm full width at half maximum (FWHM) Gaussian kernel. A high-pass filter of 180s was applied to the time series data. For a complete description of the data acquisition and preprocessing, see Woo et al. (2015).

3. Causal effects for brain regions. Observation of subject $i \in \{1, \ldots, n\}$ in run $r \in \{1, \ldots, R\}$ begins at subject specific time k_{ir}. At each equally spaced time point $t \in \{1, \ldots, T\}$ of run r, subjects are assigned no stimulus ($j = 0$) or a stimulus $j \in \{1, \ldots, J\}$. Let $z_{jtr} = 1$ if stimulus $j \in \{1, \ldots, J\}$ is applied at time point t of run r, 0; otherwise, $\mathbf{z}_{tr} \equiv (z_{1tr}, \ldots, z_{Jtr})$ the assignment vector at time t of run r, $\bar{\mathbf{z}}_{Tr} = (\mathbf{z}_{1r}, \ldots, \mathbf{z}_{Tr})$ the treatment regimen for run r of the experiment. Let $Y_{ivb,k_{ir}+t}(\bar{\mathbf{z}}) \equiv Y_{ivbtr}(\bar{\mathbf{z}})$ denote i's potential BOLD response at voxel $v^{(b)} \in \{1, \ldots, V_b\}$ of brain region $b \in \{1, \ldots, B\}$ at time t of run r under the experimental regimen $\bar{\mathbf{z}} \equiv (\bar{\mathbf{z}}_{T1}, \ldots, \bar{\mathbf{z}}_{TR})$. SL considered the case $R = 1$.They assumed responses at time t do not depend on treatments administered after time t: thus, $Y_{ivbtr}(\bar{\mathbf{z}}) = Y_{ivbtr}(\bar{\mathbf{z}}_{T1}, \ldots, \bar{\mathbf{z}}_{T,r-1}, \bar{\mathbf{z}}_{tr})$, where $\bar{\mathbf{z}}_{tr} \equiv (\mathbf{z}_{1r}, \ldots, \mathbf{z}_{tr})$. Further, we assume that responses during run r do not carry over to subsequent runs: $Y_{ivbtr}(\bar{\mathbf{z}}_{T1}, \ldots, \bar{\mathbf{z}}_{T,r-1}, \bar{\mathbf{z}}_{tr}) = Y_{ivbtr}(\bar{\mathbf{z}}_{tr})$. This is reasonable because: (a) the length of the break between runs exceeds the duration of

[1] Statistical Parametric Mapping, version 8; http://www.fil.ion.ucl.ac.uk/spm/.

the HRF, (b) unlike a problem solving task, in which a subject might continue to focus on the prior stimulus during the next run, there is no reason to think or evidence to suggest the duration of the response to the thermal stimulus exceeds the duration of the HRF, (c) the break allows the subject to recoup, mitigating potential effects due to habituation and/or fatigue.

Following SL, we decompose the potential responses as follows:

$$Y_{ivbtr}(\bar{\mathbf{z}}_{tr}) = \Psi_{ivbtr}(\bar{\mathbf{z}}_{tr}) + B_{ivbtr}(\bar{\mathbf{z}}_{tr}) + \varepsilon_{ivbtr}(\bar{\mathbf{z}}_{tr}), \tag{1}$$

where $\Psi_{ivbtr}(\bar{\mathbf{z}}_{tr})$ and $B_{ivbtr}(\bar{\mathbf{z}}_{tr})$ are, respectively, the true signal and systematic error of subject i at voxel $v^{(b)}$ during time t of run r, and $\varepsilon_{ivbtr}(\bar{\mathbf{z}}_{tr})$ is a mean *zero* error.

The signal $\Psi_{ivbt'r}(\bar{\mathbf{z}}_{tr}, \bar{\mathbf{0}}_{t'-t})$, where $\bar{\mathbf{0}}_{t'-t}$ is a vector of 0's of length $J(t'-t)$ is subject i's hemodynamic response at time $t' \geq t$ to treatments administered through time t: at each time $t'' \leq t$, a treatment $j \in \{1, \ldots, J\}$ is either administered or not, and, when treatment $j \neq 0$ is administered at time t'', a subject, treatment and voxel specific hemodynamic response function (HRF) $h_{ivbj}(q)$, $0 \leq q \leq P$, with the integer P corresponding to 30 seconds, is generated. Although the HRF varies with subjects, voxels and stimuli, its qualitative features are similar: starting from baseline A_{ivb}, initially blood flow to the voxel increases monotonically, typically peaking between four and six seconds, followed by a monotonic decrease that "overshoots" the baseline and a subsequent return to baseline. For an illustration, see Figure 1. The signal $\Psi_{ivbt'r}(\bar{\mathbf{z}}_{tr}, \bar{\mathbf{0}}_{t'-t})$ is then the sum of the baseline response A_{ivb} with the convolution of the component HRFs with treatment assignments

$$\Psi_{ivbt'r}(\bar{\mathbf{z}}_{tr}, \bar{\mathbf{0}}_{t'-t}) = A_{ivb} + \sum_{j=1}^{J} \sum_{p=t'-t}^{P} h_{ivbj}(p) z_{j,t'-p,r}. \tag{2}$$

The assumption that the relationship between neuronal activity and the HRF can be described as a linear system is often made in fMRI analysis (Lindquist (2008)). Studies have shown this assumption is reasonable (Boynton et al. (1996)), particularly if stimuli are spaced at least five seconds apart (Miezin et al. (2000)). Futher, the HRF $h_{ivbj}(\cdot)$ is assumed invariant over the course of the experiment; this is certainly reasonable when the experiment takes place within a single session. Thus, the signal (2) depends on the run r only through the sequence $\bar{\mathbf{z}}_{tr}$. Hereafter, we make this explicit: $\Psi_{ivbt'r}(\bar{\mathbf{z}}_{tr}, \bar{\mathbf{0}}_{t'-t}) \equiv \Psi_{ivbt'}(\bar{\mathbf{z}}_{tr}, \bar{\mathbf{0}}_{t'-t})$.

Differences in BOLD responses under different regimens are due to differences in the signal (causal), random errors and systematic errors of (1). Both the systematic error $B_{ivbtr}(\bar{\mathbf{z}}_{tr})$ and signal $\Psi_{ivbt}(\bar{\mathbf{z}}_{tr})$ depend on treatment regimen, but, as differences in systematic errors under different regimens are not indicative of causation, causal effects should be defined so as to exclude these. Systematic error results from machine drift and task related head motion not corrected for during preprocessing, while the *zero* mean random errors reflect measurement error due to nonneural physiological artifacts such as heart rate and respiration.

SL defined the "voxelwise unit effect" comparing treatment subregimen $\bar{\mathbf{z}}_{tr}$ with subregimen $\bar{\mathbf{z}}^*_{tr}$ for subject i at voxel $v^{(b)}$ at time $t' \geq t$ of run r as

$$\begin{aligned} \psi_{ivbt'r}(\bar{\mathbf{z}}_{tr}, \bar{\mathbf{z}}^*_{tr}) &\equiv \Psi_{ivbt'}(\bar{\mathbf{z}}_{tr}, \bar{\mathbf{0}}_{t'-t}) - \Psi_{ivbt'}(\bar{\mathbf{z}}^*_{tr}, \bar{\mathbf{0}}_{t'-t}) \equiv \psi_{ivbt'}(\bar{\mathbf{z}}_{tr}, \bar{\mathbf{z}}^*_{tr}) \\ &= \sum_{j=1}^{J} \sum_{p=t'-t}^{P} h_{ivbj}(p)\big(z_{j,t'-p,r} - z^*_{j,t'-p,r}\big). \end{aligned} \tag{3}$$

Consider now the special case where the subregimens $\bar{\mathbf{z}}_{tr}$ and $\bar{\mathbf{z}}^*_{tr}$ are identical for m or more times prior to t, and, at time t, $z_{jtr} = 1$, $z^*_{jtr} = 0$ for all j. Then, as the HRF returns to $h_{ivbj}(0) = 0$ after P time points, and $z_{j,t+1,r} = z^*_{j,t+1,r}, \ldots, z_{jt'r} = z^*_{jt'r}$, at time $t' = t + p$, $\psi_{ivbt'}(\bar{\mathbf{z}}_{tr}, \bar{\mathbf{z}}^*_{tr}) = h_{ivbj}(p)$ for $m + t' - t \geq P$.

The average effect at voxel $v^{(b)}$ at time t' of run r is the average of the unit effects $\psi_{ivbt'}(\bar{\mathbf{z}}_{tr}, \bar{\mathbf{z}}^*_{tr})$ over the population $\mathcal{P}$ from which the subjects are drawn. The variance of the unit effects may also be considered.

Regionwise unit effects may be defined using the voxelwise effects, for example, $\max_{v^{(b)}\in b}(\psi_{i1bt'}(\bar{\mathbf{z}}_{tr}, \bar{\mathbf{z}}^*_{tr}), \ldots, \psi_{iV_bbt'}(\bar{\mathbf{z}}_{tr}, \bar{\mathbf{z}}^*_{tr}))$, or $\psi_{i+bt'}(\bar{\mathbf{z}}_{tr}, \bar{\mathbf{z}}^*_{tr}) = \sum_{v^{(b)}\in b} w_{v^{(b)}} \psi_{ivbt'}(\bar{\mathbf{z}}_{tr}, \bar{\mathbf{z}}^*_{tr})$, where $0 \le w_{v^{(b)}} \le 1$ and $\sum_{v^{(b)}\in b} w_{v^{(b)}} = 1$, as here. In our application, $w_{v^{(b)}} = V_b^{-1}$, as the voxel elements have equal volume, the unit effect of subregimen $\bar{\mathbf{z}}_{tr}$ vs. $\bar{\mathbf{z}}^*_{tr}$ for subject i in region b at time t' is then

$$\begin{aligned}\psi_{i+bt'}(\bar{\mathbf{z}}_{tr}, \bar{\mathbf{z}}^*_{tr}) &= \frac{\sum_{v^{(b)}\in b}\sum_{j=1}^{J}\sum_{p=t'-t}^{P} h_{ivbj}(p)(z_{j,t'-p,r} - z^*_{j,t'-p,r})}{V_b} \\ &\equiv \sum_{j=1}^{J}\sum_{p=t'-t}^{P} h_{i+bj}(p)(z_{j,t'-p,r} - z^*_{j,t'-p,r}).\end{aligned} \tag{4}$$

The average regionwise effect of $\bar{\mathbf{z}}_{tr}$ vs. $\bar{\mathbf{z}}^*_{tr}$ in region b at time t' of run r over the population of subjects $\mathcal{P}$ is then defined as

$$\psi_{++bt'}(\bar{\mathbf{z}}_{tr}, \bar{\mathbf{z}}^*_{tr}) = E(\psi_{i+bt'}(\bar{\mathbf{z}}_{tr}, \bar{\mathbf{z}}^*_{tr})) \equiv \sum_{j=1}^{J}\sum_{p=t'-t}^{P} h_{++bj}(p)(z_{j,t'-p,r} - z^*_{j,t'-p,r}). \tag{5}$$

Let $\tilde{t} \ge t$. The variance of the regionwise unit effects and the association between these in different regions and/or time points will also be of interest,

$$\begin{aligned}&C(\psi_{i+bt'}(\bar{\mathbf{z}}_{tr}, \bar{\mathbf{z}}^*_{tr}), \psi_{i+b'\tilde{t}}(\bar{\mathbf{z}}_{tr}, \bar{\mathbf{z}}^*_{tr})) \\ &\quad = E((\psi_{i+bt'}(\bar{\mathbf{z}}_{tr}, \bar{\mathbf{z}}^*_{tr}) - \psi_{++bt'}(\bar{\mathbf{z}}_{tr}, \bar{\mathbf{z}}^*_{tr}))(\psi_{i+b'\tilde{t}}(\bar{\mathbf{z}}_{tr}, \bar{\mathbf{z}}^*_{tr}) - \psi_{++b'\tilde{t}}(\bar{\mathbf{z}}_{tr}, \bar{\mathbf{z}}^*_{tr}))).\end{aligned} \tag{6}$$

The variance measures effect heterogeneity. The standardized covariances measure the autocorrelation between responses within a region at different times or the cross-correlation between responses in different regions, the latter a measure of task related functional connectivity.

The HRFs $h_{i+bj}(\cdot)$ measure the time course of a subject's response to a single stimulus. As the building blocks underlying the causal comparisons (4), these effects are of fundamental interest, as are comparisons of these among treatments, regions and subjects. In our analysis, where interest centers on the effect of administering a stimulus j in region b during subintervals of $[0, P]$, we define unit integrated and average integrated effects between times q^* and $q^{**} > q^*$ as

$$H_{i+bj}(q^*, q^{**}) = \int_{q^*}^{q^{**}} h_{i+bj}(q)\,dq, \tag{7}$$

$$H_{++bj}(q^*, q^{**}) = \int_{q^*}^{q^{**}} h_{++bj}(q)\,dq, \tag{8}$$

respectively. Beauchamp et al. (2003) used a similar summary to capture the poststimulus increase in the hemodynamic response, prior to the subsequent undershoot.

4. Causal inference for brain regions. To estimate the effects above, additional assumptions are needed. We first construct a whole brain causal model for the decomposition (1) of the BOLD responses, then express the effects using the model parameters. An important feature of the model is the use of basis functions to estimate the signal, thereby reducing the chance of misspecification relative to specifications using the CHRF; even so, it is important to remember that, if the model is misspecified, the resulting estimates will be biased for the effects defined in Section 3. Second, we discuss the identification of the model from the observed data. Third, we discuss estimation and inference.

4.1. *A whole brain causal model.* To estimate the components of (1), we construct a whole-brain causal model below that is a generalization of the model considered in SL,

$$
\begin{aligned}
Y_{ivbtr}(\bar{\mathbf{z}}_{tr}) &= A_{ivb} + \sum_{j=1}^{J}\sum_{p=0}^{P}\sum_{k=1}^{K} D_{ivbkj} S_k(p) z_{j,t-p,r} \\
&\quad + \sum_{\ell=1}^{L} \gamma_{ivb\ell r} N_{ivbt\ell r}(\bar{\mathbf{z}}_{tr}) + \varepsilon_{ivbtr}(\bar{\mathbf{z}}_{tr}),
\end{aligned} \tag{9}
$$

with signal

$$
\begin{aligned}
\Psi_{ivbt}(\bar{\mathbf{z}}_{tr}) &= A_{ivb} + \sum_{j=1}^{J}\sum_{p=0}^{P} h_{ivbj}(p) z_{j,t-p,r} \\
&= A_{ivb} + \sum_{j=1}^{J}\sum_{p=0}^{P}\sum_{k=1}^{K} D_{ivbkj} S_k(p) z_{j,t-p,r},
\end{aligned} \tag{10}
$$

systematic error

$$
B_{ivbtr}(\bar{\mathbf{z}}_{tr}) = \sum_{\ell=1}^{L} \gamma_{ivb\ell r} N_{ivbt\ell r}(\bar{\mathbf{z}}_{tr}), \tag{11}
$$

and error $\varepsilon_{ivbtr}(\bar{\mathbf{z}}_{tr})$, with

$$
\begin{aligned}
&E\big(\varepsilon_{ivbtr}(\bar{\mathbf{z}}_{tr}) \mid A_{ivb}, \{D_{ivbkj} : (j,k) = (1,1), \ldots, (J,K)\}, \\
&\quad \{N_{ivbt\ell r}(\bar{\mathbf{z}}_{tr}) : \ell = 1, \ldots, L\}\big) = 0.
\end{aligned} \tag{12}
$$

In (10), the HRF $h_{ivbj}(\cdot) = \sum_{k=1}^{K} D_{ivbkj} S_k(\cdot)$ is modeled using basis functions $S_k(\cdot)$, $k = 1, \ldots, K$ where D_{ivbkj} is the coefficient for the kth basis function for subject i at voxel $v^{(b)}$ with respect to the jth stimulus.

The systematic error $\sum_{\ell=1}^{L} \gamma_{ivb\ell r} N_{ivbt\ell r}(\bar{\mathbf{z}}_{tr})$ at voxel $v^{(b)}$ for subject i at time t of run r results from machine drift and task related head motion not corrected for during preprocessing; the L variables $N_{ivbt\ell r}(\bar{\mathbf{z}}_{tr})$ include covariates for capturing the baseline drift and its temporal trend and measures of head motion.

In previous work, SL considered the special case $R = 1$. In addition, following common practice, SL assumed the HRF is the product of an amplitude $\mathcal{A}_{ivbj}$ with the "canonical" HRF (CHRF) widely used in the SPM neuroimaging software

$$
h_{ivbj}(q) = \mathcal{A}_{ivbj}\tilde{h}(q) = \mathcal{A}_{ivbj}\left(\frac{q^{\alpha_1-1}\beta_1^{\alpha_1}e^{-\beta_1 q}}{\Gamma(\alpha_1)} - c\frac{q^{\alpha_2-1}\beta_2^{\alpha_2}e^{-\beta_2 q}}{\Gamma(\alpha_2)}\right), \tag{13}
$$

where α_1, α_2, β_1, β_2 and c are known constants and q is measured in seconds. Because the CHRF does not depend on i, j, or $v^{(b)}$, the amplitudes $\mathcal{A}_{ivbj}$ are comparable across subjects, treatments and voxels. However, the assumption (13) is not reasonable, and its use leads to biased estimates of the HRF.

The causal estimands in Section 3 are readily expressed in terms of the model (9), for example,

$$
\begin{aligned}
\psi_{i+bt'}(\bar{\mathbf{z}}_{tr}, \bar{\mathbf{z}}^*_{tr}) &= \frac{\sum_{v^b \in b}\sum_{j=1}^{J}\sum_{k=1}^{K}\sum_{p=t'-t}^{P} D_{ivbkj} S_k(p)(z_{j,t'-p,r} - z^*_{j,t'-p,r})}{V_b} \\
&\equiv \sum_{j=1}^{J}\sum_{k=1}^{K}\sum_{p=t'-t}^{P} D_{ibkj} S_k(p)(z_{j,t'-p,r} - z^*_{j,t'-p,r})
\end{aligned} \tag{14}
$$

for $t' \le t + P$, 0 otherwise, and

$$
\begin{aligned}
H_{++bj}(q^*, q^{**}) &= \int_{q^*}^{q^{**}} h_{++bj}(q)\,dq = \int_{q^*}^{q^{**}} \sum_{k=1}^{K} \delta_{bkj} S_k(q)\,dq \\
&= \sum_{k=1}^{K} \delta_{bkj} \int_{q^*}^{q^{**}} S_k(q)\,dq,
\end{aligned} \tag{15}
$$

where δ_{bkj} is the expectation of D_{ibkj} over subjects.

4.2. *Identification of causal effects*. Let Ω denote the set of treatment regimens to which i can be exposed with positive probability and $\bar{\mathbf{Z}}_i$ the regimen to which i is assigned. In fMRI studies, in general, either all subjects are: (1) assigned to a regimen $\bar{\mathbf{z}}$, (2) randomly assigned to a regimen in Ω or (3) assigned to treatments sequentially, with later assignments depending only on earlier assignments. Let $\mathbf{Q}_i(\bar{\mathbf{z}}) = \{\Psi_{ivbt}(\bar{z}_{tr}), B_{ivbtr}(\bar{z}_{tr}), \varepsilon_{ivbtr}(\bar{z}_{tr}) : v^b \in b, b \in \{1, \ldots, B\}, (t, r) \in \{(1, 1), \ldots, (T, R)\}\}$. Then, for all $\bar{\mathbf{z}} \in \Omega$,

$$
\mathbf{Q}_i(\bar{\mathbf{z}}) \perp\!\!\!\perp \bar{\mathbf{Z}}_i, \tag{16}
$$

which implies the model (9) to (12) is identified through the analogous model for the observed data

$$
\begin{aligned}
Y_{ivbtr}(\bar{\mathbf{Z}}_{itr}) &= A_{ivb} + \sum_{j=1}^{J} \sum_{p=0}^{P} \sum_{k=1}^{K} D_{ivbkj} S_k(p) Z_{ij,t-p,r} \\
&\quad + \sum_{\ell=1}^{L} \gamma_{ivb\ell r} N_{ivbt\ell r}(\bar{\mathbf{Z}}_{itr}) + \epsilon_{ivbtr}(\bar{\mathbf{Z}}_{itr}),
\end{aligned} \tag{17}
$$

$$
\begin{aligned}
&E\big(\epsilon_{ivbtr}(\bar{\mathbf{Z}}_{itr}) \mid A_{ivb}, \big\{D_{ivbkj} : (j, k) = (1, 1), \ldots, (J, K)\big\}, \\
&\quad \big\{N_{ivbt\ell r}(\bar{\mathbf{Z}}_{itr}) : \ell = 1, \ldots L\big\}, \bar{\mathbf{Z}}_{itr}\big) = 0,
\end{aligned} \tag{18}
$$

where $\bar{\mathbf{Z}}_{itr}$ is the subregimen of $\mathbf{Z}_i$ through time t of run r and $Z_{ij,t-p,r} = 1$ if i is assigned to treatment $j \in \{1, \ldots, J\}$ at time $t - p$ of run r, 0 otherwise.

4.3. *Estimation and inference*. We now consider the observed data model (17) to (18). If the variance structure for the random effects and errors is specified, feasible generalized least squares (or maximum likelihood) estimation of the model is conceptually straightforward. But even were it possible to formulate a realistic covariance structure for the spatial relationships among voxels, the massive number of data points, parameters and random effects renders this approach infeasible. Similarly, ordinary least squares (OLS) with a robust covariance matrix is not feasible.

We therefore proceed as follows (for details, see Appendix B). The individual and average effects depend on the fixed effects δ_{bkj} and random effects $D_{ibkj} \equiv V_b^{-1} \sum_{v^{(b)} \in b} D_{ivbkj} \equiv \delta_{bkj} + d_{ibkj}$, where $E(d_{ibkj}) = 0$. For each subject, we treat the D_{ibkj} as parameters and estimate these using OLS. These estimates can be obtained using the B averaged BOLD responses $Y_{i+btr}(\bar{Z}_{itr}) \equiv V_b^{-1} \sum_{v^{(b)} \in b} Y_{ivbtr}(\bar{Z}_{itr})$ as outcomes in the aggregated model implied by (17) to (18)

$$
\begin{aligned}
Y_{i+btr}(\bar{\mathbf{Z}}_{itr}) &= A_{ib} + \sum_{j=1}^{J} \sum_{p=0}^{P} \sum_{k=1}^{K} D_{ibkj} S_k(p) Z_{ij,t-p,r} \\
&\quad + \sum_{\ell=1}^{L} \gamma_{ib\ell r} N_{i+bt\ell r}(\bar{\mathbf{Z}}_{itr}) + \epsilon_{ibtr}(\bar{\mathbf{Z}}_{itr}),
\end{aligned} \tag{19}
$$

where $N_{i+bt\ell r}(\bar{\mathbf{Z}}_{itr}) = N_{ivbt\ell r}(\bar{\mathbf{Z}}_{itr})$ for all $v^{(b)} \in b$, $A_{ib} \equiv V_b^{-1} \sum_{v^{(b)} \in b} A_{ivb} \equiv \alpha_b + a_{ib}$, $E(a_{ib}) = 0$, $\gamma_{ib\ell r} \equiv V_b^{-1} \sum_{v^{(b)} \in b} \gamma_{ivb\ell r}$ and $\epsilon_{ibtr}(\bar{\mathbf{Z}}_{itr}) = V_b^{-1} \sum_{v^{(b)} \in b} \epsilon_{ivbtr}(\bar{\mathbf{Z}}_{itr})$.

We collect the A_{ib} and D_{ibkj} together as a vector $\boldsymbol{\beta}_{i.1}$, with OLS estimator $\hat{\boldsymbol{\beta}}_{i.1} = (\hat{\boldsymbol{\beta}}'_{i11}, \ldots, \hat{\boldsymbol{\beta}}'_{iB1})'$, consisting of components $\hat{\boldsymbol{\beta}}_{ib1} = (\hat{A}_{ib}, \hat{D}_{ib11}, \ldots, \hat{D}_{ibK1}, \ldots, \hat{D}_{ibKJ})'$, $b = 1, \ldots, B$.

Using the "global two-stage method" (Davidian and Giltinan (1995)), we model $\hat{\boldsymbol{\beta}}_{i.1}$,

$$\hat{\boldsymbol{\beta}}_{i.1} = \boldsymbol{\beta}_{..1} + \mathbf{b}_{i.1} + \boldsymbol{\eta}_{i.1}, \tag{20}$$

where $\boldsymbol{\beta}_{..1} = E(\boldsymbol{\beta}_{i.1})$, $\mathbf{b}_{i.1} = \boldsymbol{\beta}_{i.1} - \boldsymbol{\beta}_{..1}$ and $\boldsymbol{\eta}_{i.1} = \hat{\boldsymbol{\beta}}_{i.1} - \boldsymbol{\beta}_{i.1}$ are independent and $\mathbf{b}_{i.1} \sim N(\mathbf{0}, \Sigma_b)$, $\boldsymbol{\eta}_{i.1} \sim N(\mathbf{0}, C_i)$. Maximum likelihood is used to estimate $\boldsymbol{\beta}_{..1}$ and Σ_b, and standard errors are obtained using the information matrix; as Davidian and Giltinan ((1995), page 141) point out, the estimator should "perform well" even if the normality assumptions are not met, as maximum likelihood and generalized least squares give the same estimate of $\boldsymbol{\beta}_{..1}$.

As the covariance matrix C_i of the OLS estimator of $\hat{\boldsymbol{\beta}}_{i.1}$ depends on unknown parameter values, in practice, an estimate $\hat{C}_i$ is used above. We constructed the estimate $\hat{C}_i$ of C_i as follows.

We assume the errors in (19) follow a multivariate AR(1) process,

$$\epsilon_{ibtr}(\bar{\mathbf{Z}}_{itr}) = \rho_b \epsilon_{ib,t-1,r}(\bar{\mathbf{Z}}_{i,t-1,r}) + u_{ibtr}, \tag{21}$$

with covariances $\mathrm{C}(u_{ibtr}, u_{ib't'r}) = \phi_{bb'}$ if $t = t'$, 0 otherwise.

The OLS estimates $\hat{A}_{ib}$, $\hat{D}_{ibkj}$, $\hat{\gamma}_{ib\ell r}$ are then used to compute residuals $e_{ibtr}(\bar{\mathbf{Z}}_{itr})$, $i = 1, \ldots, n$, and the residuals are used to estimate the parameters $\phi_{bb'}$ and ρ_b, $b = 1, \ldots, B$, $b' = 1, \ldots, B$:

$$\hat{\rho}_b = n^{-1} \sum_{i=1}^{n} \left\{ \sum_{r=1}^{R} \sum_{t=2}^{T} e_{ib,t-1,r}(\bar{\mathbf{Z}}_{i,t-1,r}) e_{ibtr}(\bar{\mathbf{Z}}_{itr}) / \sum_{r=1}^{R} \sum_{t=2}^{T} e^2_{ibtr}(\bar{\mathbf{Z}}_{itr}) \right\}, \tag{22}$$

$$\hat{\phi}_{bb'} = (1 - \hat{\rho}_b \hat{\rho}_{b'})(nTR)^{-1} \sum_{i=1}^{n} \sum_{r=1}^{R} \sum_{t=1}^{T} e_{ibtr}(\bar{\mathbf{Z}}_{itr}) e_{ib'tr}(\bar{\mathbf{Z}}_{itr}). \tag{23}$$

Equations (22) and (23) are then used to estimate the $TRB \times TRB$ covariance matrix Σ_ϵ of the errors, and the estimate $\hat{\Sigma}_\epsilon$ is used to obtain the estimate $\hat{C}_i$ of C_i.

Inference for the estimands of Section 4.1 is straightforward, as these are linear combinations of model terms. For example, (15) is a linear combination of the parameters δ_{bkj}, with known coefficients $c_k(q^*, q^{**}) = \int_{q^*}^{q^{**}} S_k(q)\, dq$. Thus, $\hat{H}_{++bj}(q^*, q^{**}) = \sum_{k=1}^{K} \hat{\delta}_{bkj} c_k(q^*, q^{**})$ is approximately normally distributed with mean $H_{++bj}(q^*, q^{**})$ and variance $\sum_{k'}^{K} \sum_{k=1}^{K} c_{k'}(q^*, q^{**}) c_k(q^*, q^{**}) C(\hat{\delta}_{bk'j}, \hat{\delta}_{bkj})$. Inferences about the correlation between neural activity in different regions can be made using the delta method.

Our application has 33 subjects, 1845 brain volumes per subject and 286 regions; with sufficiently fewer subjects, volumes and regions, the linear mixed model (19) may be estimated directly.

5. Results. We used a variant of the Yeo atlas (Yeo et al. (2011)) to subdivide the brain into 286 regions, with $j = 1$ for the warm nonpainful stimulus, $j = 2$ for the hot painful stimulus and $J = 3$ for the pain-reporting stimulus. Although stimulus $J = 3$ is not of interest here, it is necessary to model the effects of this stimulus to avoid biasing the estimated effects for stimuli 1 and 2. To model the HRFs corresponding to these stimuli, we used 15 cardinal B-spline basis functions of order 6 over the time period *zero* to 30 seconds. To model the systematic error, we included, for each region and run: (a) constant and linear terms to capture

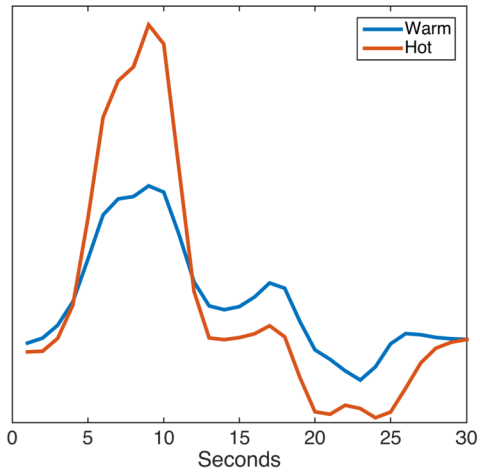

FIG. 1. *The estimated HRFs, for the warm and hot stimuli, from the anterior insula, a region commonly associated with pain processing.*

the machine drift over time, (b) the six estimated head movement parameters (x, y, z, roll, pitch and yaw) and their mean-centered squares, derivatives and squared derivatives, and (c) the signal from white matter and ventricles.

Interest centers on the neural activity associated with pain. Therefore, we compare the activity under the painful stimulus ($j = 2$) with that under the nonpainful stimulus ($j = 1$), removing from consideration activations that are the same for both conditions, due solely to the delivery of the stimuli. Figure 1 displays the estimated average HRFs ($\hat{h}_{++bj}(\cdot)$) for the anterior insula, a region strongly associated with pain affect (i.e., aversiveness of pain). As expected, while the region responds to both treatments, the signal response is higher under the painful stimulus, with a more substantial undershoot following the peak.

To more formally compare these stimuli, we performed a hypothesis test using the difference $\hat{H}_{++b2}(4, 12) - \hat{H}_{++b1}(4, 12)$ in the estimated integrated average effect from *four* to 12 seconds as a test statistic. We chose this range to cover the peak activation period of primary interest. As evidenced by Figure 1, were we to extend the range to cover the subsequent post-stimulus undershoot, we might infer (possibly correctly) that there is no difference between the painful and nonpainful stimulus, even though the HRFs are clearly different.

Results, thresholded at the 0.05 level (familywise error rate (FWER) corrected using Bonferroni correction), are shown in Figure 2. In total, 15 regions were differentially affected. There is clear activation in key lateral pain/somatosensory regions, also in the midcingulate cortex (MCC) and the dorsal lateral prefrontal cortex (DLPFC). Interestingly, we observe more cerebellar activity than in other studies using the GLM approach. Although little is known about the cerebellum's role in nociceptive processing, our results are in line with some recent work suggesting its involvement in affective processing, pain modulation, and sensorimotor processing (Moulton et al. (2010)). Altered cerebellar functioning has also been shown to be associated with chronic pain (Borsook et al. (2008)). In addition, the cerebellum

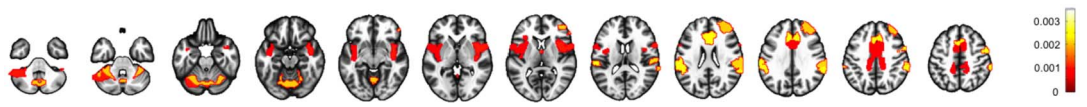

FIG. 2. *A map showing regions where the integrated average effect between four and* 12 *seconds is significantly larger in response to the hot stimulus than the warm stimulus. The results are thresholded at the* $p < 0.05$ *level (FWER corrected).*

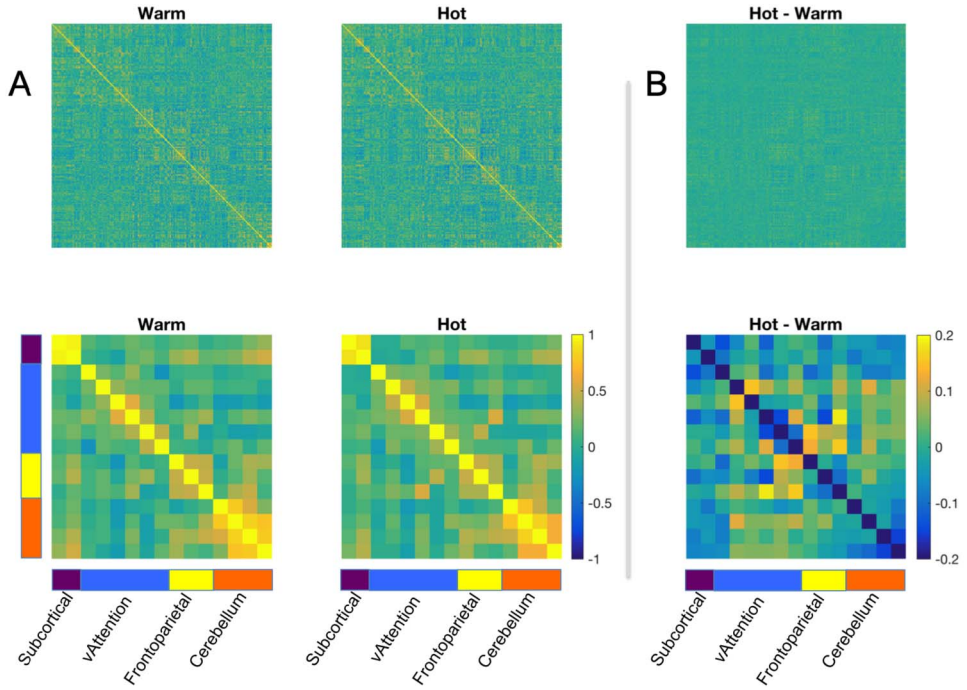

FIG. 3. (A) *The top panel displays estimates of the correlations* $\rho(H_{i+bj}(4,12), H_{i+b'j}(4,12))$ *between integrated average effects in different regions for warm* ($j = 1$) *and hot* ($j = 2$) *stimuli. The bottom panel shows the same correlations for the* 15 *differentially affected regions. The regions are grouped according to their location in one of seven networks. Here, we see subcortical regions, as well as those contained in the ventral attention* (*vAttention*) *network, frontoparietal network and cerebellum.* (B) *The difference between the estimated correlations for the hot vs. warm stimulus, for all regions* (*top*) *and for the* 15 *differentially affected regions* (*bottom*).

plays an important role in pain prediction (Wager et al. (2013)). We also observed increased variation in the estimated integrated effect for the DLPFC, a region associated with executive functioning such as sustained attention and working memory (Barbey, Koenigs and Grafman (2013)), indicating larger interindividual differences.

The top panel of Figure 3(A) displays, for all 286 regions, estimates of the correlations $\rho(H_{i+bj}(4,12), H_{i+b'j}(4,12))$ between integrated average effects in different regions, for both warm and hot stimuli. The bottom panel displays these correlations for the 15 regions previously identified as differentially affected; these are grouped into networks as defined by Yeo et al. (2011). In particular, regions were contained in the ventral attention and frontoparietal networks as well as in the cerebellum. The frontoparietal network has been shown to predict modulation of pain (Kong et al. (2013)). The ventral attention network is used when detecting sensory events outside the current focus of attention (Corbetta and Shulman (2002)). Figure 3(B) displays the difference between correlations for the two stimuli; the correlations between cerebellum and frontoparietal networks (indicated by pairs with orange color) are larger for the painful stimulus, while the correlations within the cerebellum are larger for the nonpainful stimulus (indicated by pairs colored blue).

Figure 4(A) displays the spatiotemporal correlation structure for both painful and nonpainful stimuli for the 15 significant regions. The data are organized in $15 \times 15 = 225$ blocks corresponding to each pair (b, b') for the 15 regions. Each block of dimension 31×31 displays estimates of the lagged correlations $\rho(h_{i+bj}(p), h_{i+b'j}(p'))$ between the HRFs for the equally spaced time points p, p'. As above, regions are grouped into networks as defined by Yeo et al. (2011). Results are similar for the painful and nonpainful stimuli—strong correla-

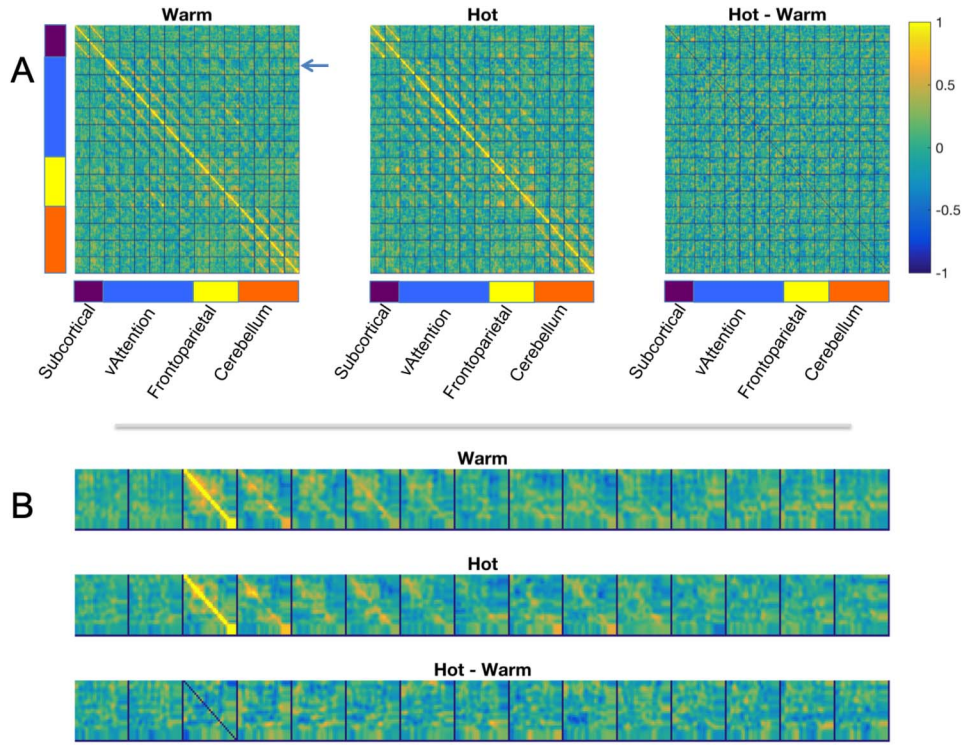

FIG. 4. (A) *Estimates of the correlations* $\rho(h_{i+bj}(p), h_{i+b'j}(p'))$*between the HRFs, corresponding to warm* ($j = 1$) *and hot* ($j = 2$) *stimuli, across the* 15 *differentially affected regions. Each square corresponds to a* 31×31 *matrix of correlations for pairs of regions* b, b'; *here* $p = 0, \ldots, 30$. *Regions are grouped by location in one of the seven networks in Yeo et al.* (2011). *The difference between the correlations under hot and warm stimuli is displayed to the right.* (B) *The top two panels show the third row of blocks depicted in Figure* 4 (*see arrow in left panel of* (A)), *corresponding to the correlation between a region in the ventral attention network and all other significant regions. The third panel depicts the difference between hot and warm stimuli.*

tion within the cerebellum, frontoparietal and ventral attention networks as well as between the frontoparietal and ventral attention networks.

Figure 4(B) highlights the correlation between a specific "seed" region from the ventral attention network and the 14 other regions deemed significant. Thus, the first two panels correspond to the third row of blocks depicted in Figure 4(A). The third panel depicts the difference between hot and warm stimuli. Here, more subtle differences between the two stimuli can be observed—an increased correlation during the latter parts of the HRF between the seed region and regions in the frontoparietal network (see, e.g., the lower right-hand portion of the ninth block from the left). This is indicative of increased correlations between regions in the time preceding pain reporting, perhaps signaling a contribution to activity during "pain recall" (e.g., Lindquist (2012)).

6. Discussion. In fMRI studies, subjects' BOLD responses to treatments are recorded at many voxels and time points. At each voxel and time point, intrasubject comparisons of signal responses under different treatment regimens yield definitions of unit treatment effects, averaging these over subjects gives average treatment effects. We build on these voxelwise effects to define unit and average treatment effects for brain regions composed of clusters of voxels, both for time points and intervals.

In the standard GLM approach to the analysis of fMRI data, each voxel is modeled separately, and the results are stitched together in a somewhat ad-hoc manner to make inferences about neural activity in aggregates composed of adjacent voxels. The approach does not yield

estimates of treatment effects for these clusters or, more generally, in brain regions. We estimate effects for regions using a multisubject, multivoxel, multi-run whole-brain causal model with explicit region parameters; the average treatment effects are a function of these parameters. Nor does the GLM approach provide estimates of the relationship between neural activity in different brain locations. Using BOLD responses averaged over regions, we model all regions simultaneously. This allows us to estimate the associations among the effects in different regions, thereby providing measures of task specific functional connectivity.

We apply the model to estimate the effects of a painful stimulus on neural activity. In addition to the activity generated in regions typically associated with pain, we observe more cerebellar activity than observed in previous work using the GLM approach. As this approach is typically implemented using the CHRF or a variant thereof, leading to downwardly biased estimates and reduced detection of activation, as demonstrated in Appendix C, this new finding points to the potential importance of using more flexible and realistic models of the HRF, as here. We present our results using whole-brain maps of effects and spatiotemporal correlation plots that display lagged relations among brain regions.

Recall that in the study on which our results are based, subjects also reported on a visual analog scale the amount of pain they experienced in response to the pain stimuli. A natural question to ask is how this subjectively experienced pain is mediated by the neural activity we have studied. To that end, we have identified those brain regions differentially affected by the warm and hot stimuli, and it is the activity in one or more of these regions that mediates the relationship between a painful stimulus and reported pain. As a future step, we want to extend the analysis here to investigate the indirect effects of the pain stimuli on reported pain through the activity in these regions. To do so, definitions of direct and indirect effects, conditions for identification and extended estimation procedures, suited to the fMRI context, will need to be developed.

Finally, in fMRI experiments subjects are typically randomly assigned to a treatment regimen prior to intervention or assignments depend only on previous assignments, and it is reasonable to assume no interference between subjects. Our definitions of unit effects for regions and our approach are also applicable under the same conditions in areas such as climate science, environmental science and geostatistics, where it is common to observe geographical units, nested within larger regions, over time. However, here if assignments depend also on previous outcomes and/or time varying confounders, it will be necessary to use identification conditions and estimation methods from the literature on longitudinal causal inference (Robins and Hernán (2009)). In addition, these effect definitions will not carry over to the case where there is interference among units (Hudgens and Halloran (2008), Sobel (2006)), in which case various kinds of effects (e.g., direct and spillover) may be of interest. If assignments also depended on spatial confounders associated with the unit and, possibly, even other units, new identification and estimation methods would be required. While challenging, the development of a general framework for spatiotemporal causal inference would be very useful; we hope our work takes a small step in this direction.

APPENDIX A: NOTATION

The following is a brief guide to the key notational conventions used in Sections 3 and 4.

Section 3.

- $t \in 1, \ldots, T$ denotes time points at which subjects are observed. Time is nested within runs $r = 1, \ldots, R$. Thus, tr refers to time point t in run r.
- $z_{jtr} = 1$ denotes the application of stimulus $j \in 1, \ldots J$ at time t of run r; otherwise, $z_{jtr} = 0$.

- $\mathbf{z}_{tr} = (z_{1tr}, \ldots z_{Jtr})$, the treatment at time t of run r.
- $\bar{\mathbf{z}}_{tr} = (\mathbf{z}_{1r}, \ldots \mathbf{z}_{tr})$, the treatment subregimen up to time t during run r.
- $\bar{\mathbf{z}} \equiv (\bar{\mathbf{z}}_{T1}, \ldots, \bar{\mathbf{z}}_{TR})$ denotes the entire treatment sequence.
- $Y_{ivbtr}(\bar{\mathbf{z}}_{tr})$ denotes the potential blood oxygen level dependent (BOLD) response of subject i at voxel $v^{(b)}$ in region b, $v^{(b)} \in \{1, \ldots, V_b\}$, $b \in \{1, \ldots, B\}$ under treatment subregimen $\bar{\mathbf{z}}_{tr}$.
- $h_{ivbj}(p)$ denotes the hemodynamic response function (HRF) for subject i at voxel $v^{(b)}$ in brain region b, p time points after the application of stimulus j.
- $h_{i+bj}(p)$ denotes the HRF for subject i in brain region b, p time points after application of stimulus j.
- $h_{++bj}(p)$ is the mean (over subjects) HRF in brain region b, p time points after application of stimulus j.
- $\Psi_{ivbt'}(\bar{\mathbf{z}}_{tr}, \bar{\mathbf{0}}_{t'-t}) = A_{ivb} + \sum_{j=1}^{J} \sum_{p=t'-t}^{P} h_{ivbj}(p) z_{j,t'-p,r}$ is the signal component of the BOLD response $Y_{ivbt'r}(\bar{\mathbf{z}}_{tr}, \mathbf{0}_{t'-t})$ at time $t' \geq t$ of run r.
- $\psi_{ivbt'}(\bar{\mathbf{z}}_{tr}, \bar{\mathbf{z}}^*_{tr})$ denotes the effect of treatment subregimen $(\bar{\mathbf{z}}_{tr}, \mathbf{0}_{t'-t})$ vs. $(\bar{\mathbf{z}}^*_{tr}, \mathbf{0}_{t'-t})$ for subject i at voxel $v^{(b)}$ of region b at time $t' \geq t$ of run r.
- $B_{ivbtr}(\bar{\mathbf{z}}_{tr})$ denotes the systematic error component of $Y_{ivbtr}(\bar{\mathbf{z}}_{tr})$.
- $\varepsilon_{ivbtr}(\bar{\mathbf{z}}_{tr})$ is the mean zero random error component of $Y_{ivbtr}(\bar{\mathbf{z}}_{tr})$.
- $H_{i+bj}(q^*, q^{**}) = \int_{q^*}^{q^{**}} h_{i+bj}(q)\, dq$ denotes the unit integrated effect for subject i in region b under stimulus j from time q^* to q^{**}.
- $H_{++bj}(q^*, q^{**}) = \int_{q^*}^{q^{**}} h_{++bj}(q)\, dq$ denotes the average integrated effect (over subjects) in region b, under stimulus j, from time q^* to q^{**}.

Section 4.

- $Z_{ij,t-p,r} = 1$ if stimulus j is applied to subject i at time $t - p$ of run r, 0 otherwise.
- $\bar{\mathbf{Z}}_{itr}$ denotes the observed sequence of stimuli applied to subject i through time t during run r.
- $\bar{\mathbf{Z}}_i$ denotes the treatment regimen applied to subject i.
- The signal $\Psi_{ivbtr}(\bar{\mathbf{z}}_{tr})$ is expressed using basis functions for the HRF, $h_{ivbj}(p) = A_{ivb} + \sum_{k=1}^{K} D_{ivbkj} S_k(p)$, where $S_k(\cdot)$, $k = 1, \ldots, K$ are K basis functions.
- $h_{i+bj}(p) = A_{ib} + \sum_{k=1}^{K} D_{ibkj} S_k(p)$, the HRF averaged over voxels in region b.
- $h_{++bj}(p) = E(h_{i+bj}(p)) = \alpha_b + \sum_{k=1}^{K} \delta_{bkj} S_k(p)$.

APPENDIX B

We show that the least squares estimates $\hat{A}_{ib}$ of A_{ib}, $b = 1, \ldots, B$ and $\hat{D}_{ibkj}$ of D_{ibkj}, $b = 1, \ldots, B$, $k = 1, \ldots, K$, $j = 1, \ldots, J$, in model (19) are the averages over region b of the least squares estimates $\hat{A}_{ivb}$ and $\hat{D}_{ivbkj}$ of A_{ivb} and D_{ivbkj} obtained using the GLM approach in which i's BOLD response series at each voxel is treated as a separate response vector. As it is not feasible to estimate the whole-brain model using the voxelwise BOLD responses, we therefore apply OLS to the BOLD responses averaged over regions and, then, model these estimates.

Let $\mathbf{Y}_{ivb.r} = (Y_{ivb1r}(\mathbf{Z}_{i1r}), \ldots, Y_{ivbTr}(\bar{\mathbf{Z}}_{iTr}))'$, $\mathbf{Y}_{ivb..} = (\mathbf{Y}'_{ivb.1}, \ldots, \mathbf{Y}'_{ivb.R})'$, $\mathbf{D}_{ivb.j} = (D_{ivb1j}, \ldots, D_{ivbKj})'$, $j = 1, \ldots, J$, $\boldsymbol{\beta}_{ivb1} = (A_{ivb}, \mathbf{D}'_{ivb.1}, \ldots, \mathbf{D}'_{ivb.J})'$, $X_{i1} = (\mathbf{1}, W_i)$, where $\mathbf{1}$ is $TR \times 1$ and $W_i = (W_{i1}, \ldots, W_{iJ})$ is the $TR \times KJ$ matrix composed of the $T \times K$ submatrices W_{ij} with elements $\sum_{p=0}^{P} Z_{i,j,t-p,r} S_k(p)$ in position $(T(R-1) + t, k)$ of W_{ij}. Let $\boldsymbol{\beta}_{ivb2r} = (\gamma_{ivb1r}, \ldots, \gamma_{ivbLr})'$, X_{i2r} the corresponding $T \times L$ matrix of nuisance covariates, with element $N_{ivbt\ell}$ in position (t, ℓ) $\boldsymbol{\beta}_{ivb2.} = (\boldsymbol{\beta}'_{ivb21}, \ldots, \boldsymbol{\beta}'_{ivb2R})'$,

$X_{i2.} = \text{diag}(X_{i21}, \ldots, X_{i2R})$, $\boldsymbol{\epsilon}_{ivb.r} = (\epsilon_{ivb1r}(\mathbf{Z}_{i1r}), \ldots, \epsilon_{ivbTr}(\bar{\mathbf{Z}}_{iTr}))'$, $\boldsymbol{\epsilon}_{ivb..} = (\boldsymbol{\epsilon}'_{ivb.1}, \ldots, \boldsymbol{\epsilon}'_{ivb.R})'$. For individual i, reexpressing (17) at a single voxel $v^{(b)}$ in matrix form gives

$$\mathbf{Y}_{ivb..} = X_{i1}\boldsymbol{\beta}_{ivb1} + X_{i2.}\boldsymbol{\beta}_{ivb2.} + \boldsymbol{\epsilon}_{ivb..}. \tag{24}$$

The model matrix $(X_{i1}, X_{i2.})$ is assumed to have full column rank. Some simple matrix algebra gives

$$\hat{\boldsymbol{\beta}}_{ivb1} = Q_i^{-1} G_i \mathbf{Y}_{ivb..}, \tag{25}$$

where

$$Q_i = X'_{i1}X_{i1} - (X'_{i1}X_{i2.})(X'_{i2.}X_{i2.})^{-1}(X'_{i2.}X_{i1}), \tag{26}$$

$$G_i = X'_{i1} - (X'_{i1}X_{i2.})(X'_{i2.}X_{i2.})^{-1}X'_{i2.}. \tag{27}$$

Now, let $\mathbf{Y}_{i....} = (\mathbf{Y}'_{i11..}, \ldots, \mathbf{Y}'_{iV_11..}, \ldots, \mathbf{Y}'_{iV_BB..})'$, $\mathbf{D}_{ib.j} = (D_{ib1j}, \ldots, D_{ibKj})'$, $\boldsymbol{\beta}_{ib1} = (A_{ib}, \mathbf{D}'_{ib.1}, \ldots, \mathbf{D}'_{ib.J})'$, $\boldsymbol{\beta}_{i.1} = (\boldsymbol{\beta}'_{i11}, \ldots, \boldsymbol{\beta}'_{iB1})'$. Let $V = \sum_{b=1}^{B} V_b$; let the $VTR \times B(KJ+1)$ matrix $X^*_{i1} = J \otimes X_{i1}$, where $J = (\mathbf{j}_1, \ldots, \mathbf{j}_B)$ is a $V \times B$ matrix with columns $\mathbf{j}_b = (\mathbf{0}_{1\times V_1+\ldots V_{b-1}}, \mathbf{1}_{1\times V_b}, \mathbf{0}_{1\times(V-(V_1+\cdots V_b))})'$, $b = 1, \ldots, B$ and $\otimes$ denotes the Kronecker product. Let $\boldsymbol{\beta}_{i..2.} = (\boldsymbol{\beta}'_{i112.}, \ldots, \boldsymbol{\beta}'_{iV_112.}, \ldots, \boldsymbol{\beta}'_{iV_BB2.})'$, X^*_{i2} the $VTR \times VLR$ matrix $I_V \otimes X_{i2.}$. Let $a_{ivb} = A_{ivb} - A_{ib}$, $d_{ivbkj} = D_{ivbkj} - D_{ibkj}$,

$$\eta_{ivbtr}(\bar{\mathbf{Z}}_{itr}) = a_{ivb} + \sum_{j=1}^{J}\sum_{p=0}^{P}\sum_{k=0}^{K} d_{ivbkj} Z_{ij,t-p,r} S_k(p) + \epsilon_{ivbtr}(\bar{\mathbf{Z}}_{itr}), \tag{28}$$

$\boldsymbol{\eta}_{ivb.r} = (\eta_{ivb1r}(\mathbf{Z}_{i1r}), \ldots, \eta_{ivbTr}(\bar{\mathbf{Z}}_{iTr}))'$, $\boldsymbol{\eta}_{ivb..} = (\boldsymbol{\eta}'_{ivb.1}, \ldots, \boldsymbol{\eta}'_{ivb.R})'$ and $\boldsymbol{\eta}_{i....} = (\boldsymbol{\eta}'_{i11..}, \ldots, \boldsymbol{\eta}'_{iBV_B..})'$.

The whole-brain voxel level model for individual i is

$$\mathbf{Y}_{i....} = X^*_{i1}\boldsymbol{\beta}_{i.1} + X^*_{i2}\boldsymbol{\beta}_{i..2.} + \boldsymbol{\eta}_{i....}, \tag{29}$$

with least squares estimator $\hat{\boldsymbol{\beta}}_{i.1}$ of $\boldsymbol{\beta}_{i.1}$,

$$\hat{\boldsymbol{\beta}}_{i.1} = Q^{*-1}_i G^*_i \mathbf{Y}_{i....}, \tag{30}$$

where the $B(KJ+1) \times B(KJ+1)$ matrix

$$Q^*_i = X^{*\prime}_{i1}X^*_{i1} - (X^{*\prime}_{i1}X^*_{i2})(X^{*\prime}_{i2}X^*_{i2})^{-1}(X^{*\prime}_{12}X^*_{i1}) = \text{diag}(V_1, \ldots, V_B) \otimes Q_i \tag{31}$$

and the $B(KJ+1) \times VTR$ matrix

$$G^*_i = X^{*\prime}_{i1} - (X^{*\prime}_{i1}X^*_{i2})(X^{*\prime}_{i2}X^*_{i2})^{-1}X^{*\prime}_{i2} = J' \otimes G_i, \tag{32}$$

giving

$$\hat{\boldsymbol{\beta}}_{i.1} = (\text{diag}(V_1^{-1}, \ldots, V_B^{-1}) \otimes Q_i^{-1})(J' \otimes G_i)\mathbf{Y}_{i....} = (J^{*\prime} \otimes Q_i^{-1}G_i)\mathbf{Y}_{i....}, \tag{33}$$

where row 1 of the $B \times V$ matrix $J^{*\prime} = \text{diag}(V_1^{-1}, \ldots, V_B^{-1})J'$ has entries, V_1^{-1} in columns $1, \ldots, V_1$, 0 otherwise, …, row B has entries V_B^{-1} in columns $V_{B-1}+1, \ldots, V_B$, 0 otherwise; thus, $J^{*\prime} \otimes Q_i^{-1}G_i$ consists of BV blocks of size $(KJ+1) \times TR$ and block bv (corresponding to element bv in $J^{*\prime}$) is $V_b^{-1}Q_i^{-1}G_i$ for voxels in region b, 0 otherwise. Thus, the least squares estimator $\hat{\boldsymbol{\beta}}_{ib1}$ of $\boldsymbol{\beta}_{ib1}$, $b = 1, \ldots, B$, is

$$\hat{\boldsymbol{\beta}}_{ib1} = V_b^{-1}\sum_{v^{(b)}\in b} Q_i^{-1}G_i\mathbf{Y}_{ivb..} = V_b^{-1}\sum_{v^{(b)}\in b}\hat{\boldsymbol{\beta}}_{ivb1}. \tag{34}$$

Further, as $V_b^{-1}\sum_{v^{(b)}\in b} Q_i^{-1}G_i\mathbf{Y}_{ivb..} = Q_i^{-1}G_i(V_b^{-1}\sum_{v^{(b)}\in b}\mathbf{Y}_{ivb..})$, the least squares estimates may be computed by first averaging over the voxels in the region,

$$\hat{\boldsymbol{\beta}}_{ib1} = Q_i^{-1}G_i\mathbf{Y}_{i+b..}, \tag{35}$$

where $\mathbf{Y}_{i+b..} = V_b^{-1}\sum_{v^{(b)}\in b}\mathbf{Y}_{ivb..}$.

Next, we model $\mathbf{Y}_{i+\cdots} = (\mathbf{Y}'_{i+1..},\ldots,\mathbf{Y}'_{i+B..})'$. For $b = 1,\ldots,B$

$$\mathbf{Y}_{i+b..} = X_{i1}\boldsymbol{\beta}_{ib1} + X_{i2}\boldsymbol{\beta}_{ib2} + \boldsymbol{\epsilon}_{ib..}, \tag{36}$$

where $\boldsymbol{\epsilon}_{ib.r} = (\epsilon_{ib1r},\ldots,\epsilon_{ibTr})'$, $\boldsymbol{\epsilon}_{ib..} = (\boldsymbol{\epsilon}'_{ib.1},\ldots,\boldsymbol{\epsilon}'_{ib.R})'$. Now, let $\boldsymbol{\epsilon}_{i\cdots} = (\boldsymbol{\epsilon}'_{i1..},\ldots,\boldsymbol{\epsilon}'_{iB..})'$, $X_i = (X_{i1}, X_{i2})$, $\boldsymbol{\beta}_{i..} = (\boldsymbol{\beta}'_{i11},\boldsymbol{\beta}'_{i12},\ldots,\boldsymbol{\beta}'_{iB1},\boldsymbol{\beta}'_{iB2})'$. The whole-brain region level model for individual i is

$$\mathbf{Y}_{i+\cdots} = (I_B\otimes X_i)\boldsymbol{\beta}_{i..} + \boldsymbol{\epsilon}_{i\cdots}, \tag{37}$$

where I_B is the $B\times B$ identity matrix, $\epsilon_{ibtr} = \rho_b\epsilon_{ib,t-1,r} + u_{ibtr}$, the vectors $\mathbf{u}_{i.tr} = (u_{i1tr},\ldots,u_{iBtr})'$, $t = 1,\ldots,T$, $r = 1,\ldots,R$ are independent and identically distributed $\mathrm{N}(\mathbf{0},\Phi)$. We also assume the collection of R vectors $\boldsymbol{\epsilon}_{i..r} = (\epsilon_{i11r},\ldots,\epsilon_{i1Tr},\ldots,\epsilon_{iB1r},\ldots,\epsilon_{iBTr})'$ are mutually independent. Thus, the covariance matrix

$$\Sigma_\epsilon = \mathrm{Var}(\boldsymbol{\epsilon}_{i\cdots}) = \begin{bmatrix} I_R\otimes V_{11},\ldots,I_R\otimes V_{1B} \\ \vdots \quad\quad \vdots \\ I_R\otimes V_{B1},\ldots,I_R\otimes V_{BB}\end{bmatrix}$$

consists of B^2 blocks $I_R\otimes V_{bb'}$, where $V_{bb'} = \mathrm{Cov}(\boldsymbol{\epsilon}_{ib.r},\boldsymbol{\epsilon}_{ib.'r})$ is the $T\times T$ matrix with elements $\mathrm{cov}(\varepsilon_{ibtr},\varepsilon_{i,b't't'r}) = (\phi_{bb'}/1-\rho_b\rho_{b'})\rho_b^{\max(0,t-t')}\rho_{b'}^{\max(0,t'-t)}$, $r = 1,\ldots R$. The least squares estimator $\hat{\boldsymbol{\beta}}_{i..}$ of $\boldsymbol{\beta}_{i..}$ has covariance matrix

$$V(\hat{\boldsymbol{\beta}}_{i..}) = \big(I_B\otimes X_i'X_i\big)^{-1}\big[(I_B\otimes X_i)\Sigma_\epsilon(I_B\otimes X_i)\big]\big(I_B\otimes X_i'X_i\big)^{-1}; \tag{38}$$

$V(\hat{\boldsymbol{\beta}}_{i..})$ is then estimated using (22) and (23) to estimate Σ_ϵ, with the resulting estimator $\hat{\Sigma}_\epsilon$ used in place of Σ_ϵ in (38).

Now, we model the estimates $\hat{\boldsymbol{\beta}}_{i.1} = (\hat{\boldsymbol{\beta}}'_{i11},\ldots,\hat{\boldsymbol{\beta}}'_{iB1})'$ of $\boldsymbol{\beta}_{i.1}$ using the decomposition (20). The log-likelihood

$$\ell(\boldsymbol{\beta}_{..1},\Sigma_b) \propto -\sum_{i=1}^{n}\big[\ln|\Sigma_b + C_i| + (\hat{\boldsymbol{\beta}}_{i.1}-\boldsymbol{\beta}_{..1})'(\Sigma_b+C_i)^{-1}(\hat{\boldsymbol{\beta}}_{i.1}-\boldsymbol{\beta}_{..1})\big]. \tag{39}$$

As C_i is unknown, the estimated asymptotic covariance matrix $\hat{C}_i$ is used in place of C_i to solve the likelihood equations and obtain standard errors:

$$\frac{\partial\ell}{\partial\boldsymbol{\beta}_{..1}} = \sum_{i=1}^{n}(\Sigma_b+\hat{C}_i)^{-1}(\hat{\boldsymbol{\beta}}_{i.1}-\boldsymbol{\beta}_{..1}) = \mathbf{0}, \tag{40}$$

$$\begin{aligned}\frac{\partial\ell}{\partial\sigma_{qs}} = -1/2\sum_{i=1}^{n}\Bigg[tr\bigg((\Sigma_b+\hat{C}_i)^{-1}\frac{\partial\Sigma_b}{\partial\sigma_{qs}}\bigg) \\ -(\hat{\boldsymbol{\beta}}_{1.1}-\boldsymbol{\beta}_{..1})'(\Sigma_b+\hat{C}_i)^{-1}\frac{\partial\Sigma_b}{\partial\sigma_{qs}}(\Sigma_b+\hat{C}_i)^{-1}(\hat{\boldsymbol{\beta}}_{1.1}-\boldsymbol{\beta}_{..1})\Bigg] = 0,\end{aligned} \tag{41}$$

where $\sigma_{qs} = \sigma_{sq}$, $q\le s$, is the qs element of Σ_b. Letting $(A)_{qs}$ denote the (qs) element of a matrix A, note that (41) reduces further, using $tr((\Sigma_b+\hat{C}_i)^{-1}\frac{\partial\Sigma_b}{\partial\sigma_{qs}}) = 2((\Sigma_b+\hat{C}_i)^{-1})_{qs}$ for

$q \neq s$ and $((\Sigma_b + \hat{C}_i)^{-1})_{qs}$ if $q = s$. The information matrix has components

$$-E\left(\frac{\partial^2 \ell}{\partial \boldsymbol{\beta}_{\cdot 1} \partial \boldsymbol{\beta}'_{\cdot 1}}\right) = \sum_{i=1}^{n} (\Sigma_b + \hat{C}_i)^{-1}, \tag{42}$$

$$-E\left(\frac{\partial^2 \ell}{\partial \boldsymbol{\beta}_{\cdot 1} \partial \sigma_{qs}}\right) = \mathbf{0}, \quad q \leq s, \tag{43}$$

$$-E\left(\frac{\partial^2 \ell}{\partial \sigma_{qs} \partial \sigma_{q's'}}\right) = 1/2 \sum_{i=1}^{n} tr\left[(\Sigma_b + \hat{C}_i)^{-1} \frac{\partial \Sigma_b}{\partial \sigma_{rs}} (\Sigma_b + \hat{C}_i)^{-1} \frac{\partial \Sigma_b}{\partial \sigma_{q's'}}\right] \tag{44}$$

$$= 1/2 \sum_{i=1}^{n} [((\Sigma_b + \hat{C}_i)^{-1})_{qs} ((\Sigma_b + \hat{C}_i)^{-1})_{q's'}], \tag{45}$$

where $q \leq s$, $q' \leq s'$.

The EM-algorithm is used to solve the likelihood equations (40) and (41). Here, the "E-step" provides estimates $\hat{\boldsymbol{\beta}}_{i\cdot 1}$, while the "M-step" provides estimates of the population parameters $\boldsymbol{\beta}_{\cdot 1}$ and Σ_b. The asymptotic covariance of $\hat{\boldsymbol{\beta}}_{\cdot 1}$ can be obtained by inverting the estimated information matrix with estimates $\hat{\boldsymbol{\beta}}_{\cdot 1}$ and $\hat{\Sigma}_b$ in place of $\boldsymbol{\beta}_{\cdot 1}$ and Σ_b.

APPENDIX C: DETECTING ACTIVATION—A SIMULATION STUDY

We conduct a small simulation study to compare the performance of our method for detecting and estimating activations in ROIs with that of the standard GLM approach, in which each voxel is modeled separately as the product of an amplitude with the known CHRF. Attention is limited to the case where the voxel activations are homogeneous throughout the region. The setup is similar to that in Lindquist et al. (2009) and Degras and Lindquist (2014). It is also important to note that the standard GLM approach, because each voxel is modeled separately, cannot be used to study functional connectivity between ROIs; if this is of interest, a whole-brain approach is essential.

As shown in Figure 5(A), within a static brain slice of size 51×40, a set of 25 identically sized squares, each size 4×4, were placed to represent active ROIs. In each square, a different HRF was created using stimulus functions that vary systematically across squares in terms of onset and duration. The HRF in the upper left-hand corner is the CHRF. From left to right the onset of activation varied from the first to the fifth TR, and from top to bottom, the duration of activation varied from one to nine TRs in steps of two. Figure 5(B) shows the *five* HRFs with no onset shift which are representative of the remaining HRFs. The TR is *one* second and the time between stimuli was set to 30 seconds. All HRFs have an amplitude of one. This activation pattern was repeated to simulate a total of 10 trials; hence, $T = 300$ in our simulation.

We generated 1000 datasets, each consisting of the BOLD responses of 15 subjects. Equation (17) with square specific HRFs was used to to generate the responses, assuming $A_{ivb} = 0$, $J = 1$, $R = 1$, and no systematic error. An event-related stimulus function with a single spike repeated every 30 seconds was used. Within each square, a subject specific amplitude D_{ivb1}, equal for all voxels $v^{(b)} \in b$, was drawn from the normal distribution $\mathcal{N}(1, 4/3)$ with mean 1, variance $4/3$ and an error $\epsilon_{ivbt1}(\bar{Z}_{it1})$ was drawn from the $\mathcal{N}(0, 4)$ distribution. This setup yields an effect size (Cohen's d = 0.5), similar to that observed in the visual and motor cortex (Wager et al. (2005)).

For each dataset, we used: (1) the standard approach to estimate the HRF as the product of the CHRF with an estimated amplitude and (2) our approach, in which the HRF in every square is estimated using 15 B-spline basis functions of order 6. We parceled the brain slice

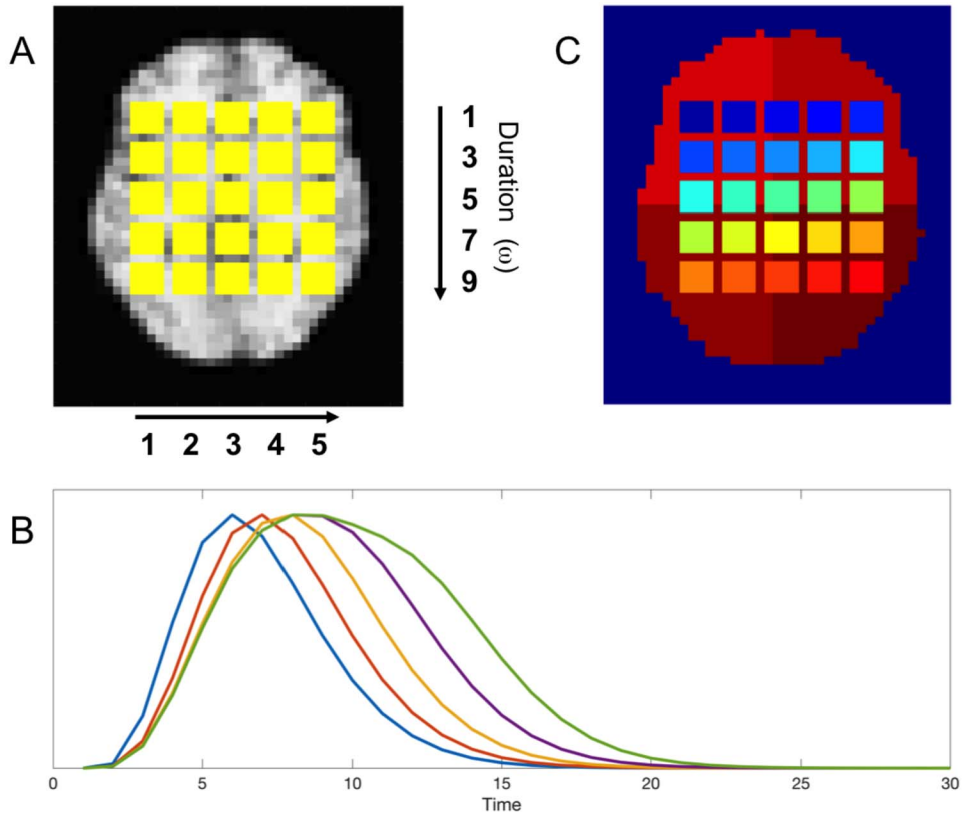

FIG. 5. *Overview of the simulation.* (A) *A set of* 25 *equally sized squares were placed within a static brain slice to represent regions of interest* (*ROIs*). *The HRFs vary systematically across the squares in their onset and duration of neuronal activation. From left to right the onset of activation varied between the squares from the first to the fifth TR. From top to bottom, the duration of activation varied from one to nine TR in steps of two.* (B) *Five HRFs with identical onset and varying duration.* (C) *The parcellation scheme used to define ROIs for our method.*

into 29 regions, as shown in Figure 5(C), corresponding to the 25 activation profiles and four inactive background regions.

Using our approach and the integrated average effect (15) between $q^* = 4$ s and $q^{**} = 12$ s as a test statistic, we: (i) estimated the bias of the test statistic at each voxel and (ii) tested the null hypothesis of no activation: $H_0 : H_{++b1}(q^*, q^{**}) = 0$ vs. $H_A : H_{++b1}(q^*, q^{**}) > 0$; the range [4, 12] was chosen to cover the peak activation period of primary interest. To control for multiple comparisons, we used the Benjami–Hochberg (Benjamini and Hochberg (1995)) procedure with the false discovery rate set at 0.05. The standard GLM approach estimates an amplitude for each subject and voxel, using these to estimate the population amplitude and its variance. Since the HRF is assumed to be the product of the CHRF with the amplitude, testing for activation under this approach is equivalent to testing if the population amplitude is 0, that is, $H_0 : D_{+vb1} = 0$ vs. $H_A : D_{+vb1} > 0$, where $D_{+vb1} = E(D_{ivb1})$.

The top row of Figure 6 displays the mean bias for the integrated average effect over the 1000 repetitions for each voxel in the slice. Clearly, the GLM approach provides an unbiased estimate in the upper left-hand corner where the HRF is correctly specified. However, its performance worsens dramatically as the onset and/or duration of the HRF increases. Interestingly, the bias is consistently negative, indicating that the GLM will consistently underestimate the average integrated effect. Using our approach, the bias is more than an order of magnitude smaller. In addition, there is no apparent spatial pattern for the bias, and it takes both positive and negative values.

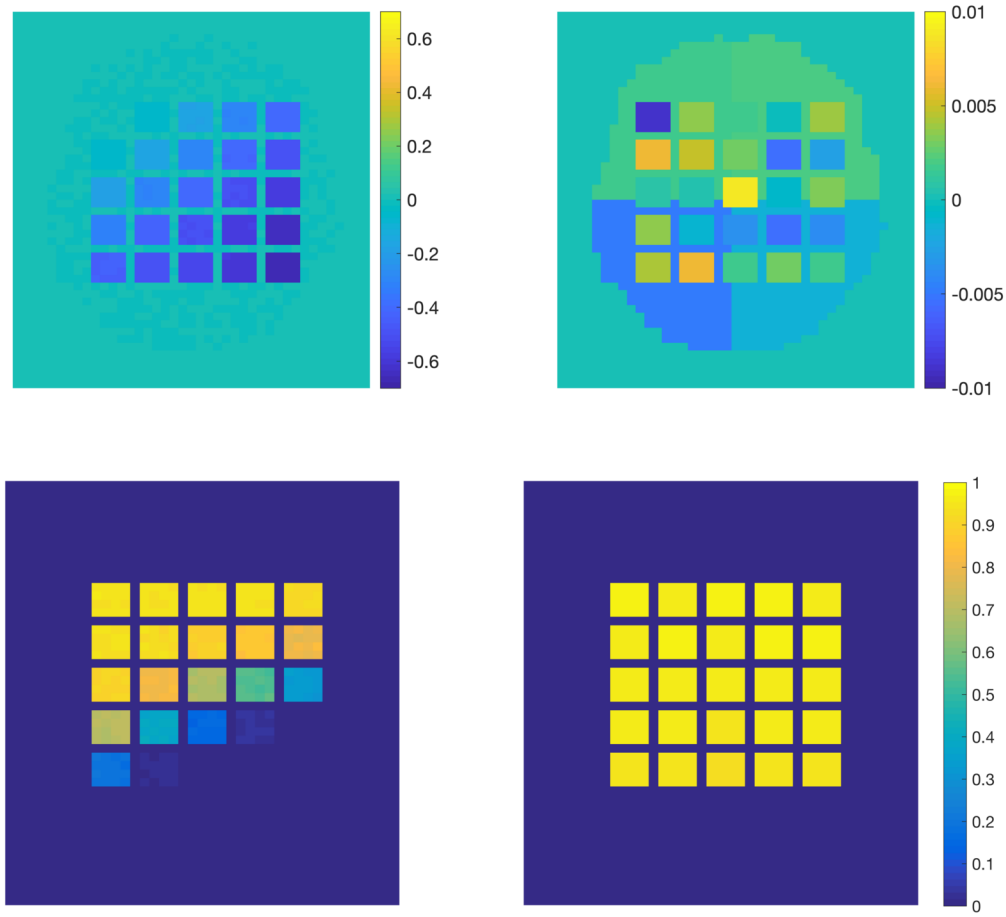

FIG. 6. (*Top row*) *The mean bias for the integrated average effect over the* 1000 *replications of the simulation for each voxel using the standard voxel-wise GLM* (*left*) *approach and our proposed approach* (*right*). *For the GLM approach, the bias is roughly zero in the upper left-hand corner where the HRF is correct, and increases as the HRF begins to differ from its canonical form. Note that the bias is consistently negative. For our approach the bias is at least an order of magnitude smaller with no consistent spatial pattern.* (*Bottom row*) *The proportion of times in* 1000 *replications of the simulation that each voxel was deemed active using the standard voxel-wise GLM* (*left*) *approach and our proposed approach* (*right*). *Note that since our approach is fit at the region-level all voxels within the same region will have the same proportion. For our approach the average proportion in the active voxels is* 0.960. *Using the GLM it is roughly the same in the upper left-hand corner where the HRF is correct, and decays as the HRF begins to differ from its canonical form.*

The bottom row of Figure 6 shows the proportion of times each voxel in the slice was deemed active in the 1000 repetitions. The GLM approach gives reasonable results for delayed onsets within *three* seconds and durations up to *three* seconds, corresponding to squares in the upper left-hand corner. However, its performance worsens dramatically as onset and duration increase; for example, in the square in upper right-hand corner the proportion deemed active is 0.91, in the lower left-hand corner it is 0.19 and in the lower right-hand corner it is 0. The proportions of false positives are well controlled in the background voxels, as the average proportion deemed active is only 0.0014, indicating a high degree of specificity.

Using our approach, irrespective of the shape of the underlying HRF we recover appropriate activations. The average proportion of true positives for the integrated average effect across the 25 squares is 0.96, indicating high sensitivity. The average proportion deemed active in the background regions is 0.001, illustrating the method's specificity.

Acknowledgments. We thank Tor Wager for supplying the data. For helpful comments, we thank the anonymous reviewers and the Editor in charge of the manuscript. The code and data is available on the authors GitHub page (https://github.com/mal2053/CausalCode/). It consists of MATLAB code implementing the methods from the paper.

This research was supported by NIH Grant R01EB016061.

REFERENCES

BARBEY, A. K., KOENIGS, M. and GRAFMAN, J. (2013). Dorsolateral prefrontal contributions to human working memory. *Cortex* **49** 1195–1205.

BEAUCHAMP, M. S., LEE, K. E., HAXBY, J. V. and MARTIN, A. (2003). FMRI responses to video and point-light displays of moving humans and manipulable objects. *J. Cogn. Neurosci.* **15** 991–1001.

BENJAMINI, Y. and HOCHBERG, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *J. Roy. Statist. Soc. Ser. B* **57** 289–300. MR1325392

BORSOOK, D., MOULTON, E., TULLY, S., SCHMAHMANN, J. and BECERRA, L. (2008). Human cerebellar responses to brush and heat stimuli in healthy and neuropathic pain subjects. *Cerebellum* **7** 252–272.

BOWMAN, F. D. (2005). Spatio-temporal modeling of localized brain activity. *Biostatistics* **6** 558–575.

BOWMAN, F. D. (2007). Spatiotemporal models for region of interest analyses of functional neuroimaging data. *J. Amer. Statist. Assoc.* **102** 442–453. MR2370845 https://doi.org/10.1198/016214506000001347

BOWMAN, F. D., CAFFO, B., BASSETT, S. S. and KILTS, C. (2008). A Bayesian hierarchical framework for spatial modeling of fMRI data. *NeuroImage* **39** 146–156.

BOYNTON, G. M., ENGEL, S. A., GLOVER, G. H. and HEEGER, D. J. (1996). Linear systems analysis of functional magnetic resonance imaging in human V1. *J. Neurosci.* **16** 4207–4221.

BUXTON, R. B. (2009). *Introduction to Functional Magnetic Resonance Imaging*: *Principles and Techniques*. Cambridge Univ. Press, Cambridge.

CARP, J. (2012). On the plurality of (methodological) worlds: Estimating the analytic flexibility of FMRI experiments. *Front. Neurosci.* **6** 149. https://doi.org/10.3389/fnins.2012.00149

CORBETTA, M. and SHULMAN, G. L. (2002). Control of goal-directed and stimulus-driven attention in the brain. *Nat. Rev., Neurosci.* **3** 201–215. https://doi.org/10.1038/nrn755

DAVIDIAN, M. and GILTINAN, D. M. (1995). *Nonlinear Models for Repeated Measurement Data*, 1st ed. Chapman and Hall/CRC, New York.

DEGRAS, D. and LINDQUIST, M. A. (2014). A hierarchical model for simultaneous detection and estimation in multi-subject fMRI studies. *NeuroImage* **98** 61–72. https://doi.org/10.1016/j.neuroimage.2014.04.052

FRISTON, K. J. (2011). Functional and effective connectivity: A review. *Brain Connect.* **1** 13–36. https://doi.org/10.1089/brain.2011.0008

FRISTON, K. J., HARRISON, L. and PENNY, W. (2003). Dynamic causal modelling. *NeuroImage* **19** 1273–1302.

FRISTON, K. J., HOLMES, A. P., WORSLEY, K. J., POLINE, J.-P., FRITH, C. D. and FRACKOWIAK, R. S. (1994). Statistical parametric maps in functional imaging: A general linear approach. *Hum. Brain Mapp.* **2** 189–210.

HARRISON, L. M. and GREEN, G. G. (2010). A Bayesian spatiotemporal model for very large data sets. *NeuroImage* **50** 1126–1141.

HUDGENS, M. G. and HALLORAN, M. E. (2008). Toward causal inference with interference. *J. Amer. Statist. Assoc.* **103** 832–842. MR2435472 https://doi.org/10.1198/016214508000000292

KONG, J., JENSEN, K., LOIOTILE, R., CHEETHAM, A., WEY, H.-Y., TAN, Y., ROSEN, B., SMOLLER, J. W., KAPTCHUK, T. J. et al. (2013). Functional connectivity of the frontoparietal network predicts cognitive modulation of pain. *Pain* **154** 459–467.

KWONG, K. K., BELLIVEAU, J. W., CHESLER, D. A., GOLDBERG, I. E., WEISSKOFF, R. M., PONCELET, B. P., KENNEDY, D. N., HOPPEL, B. E., COHEN, M. S. et al. (1992). Dynamic magnetic resonance imaging of human brain activity during primary sensory stimulation. *Proc. Natl. Acad. Sci. USA* **89** 5675–5679.

LINDQUIST, M. A. (2008). The statistical analysis of fMRI data. *Statist. Sci.* **23** 439–464. MR2530545 https://doi.org/10.1214/09-STS282

LINDQUIST, M. A. (2012). Functional causal mediation analysis with an application to brain connectivity. *J. Amer. Statist. Assoc.* **107** 1297–1309. MR3036396 https://doi.org/10.1080/01621459.2012.695640

LINDQUIST, M. A. and SOBEL, M. E. (2011). Graphical models, potential outcomes and causal inference: Comment on Ramsey, Spirtes and Glymour. *NeuroImage* **57** 334–336.

LINDQUIST, M. A. and SOBEL, M. E. (2013). Cloak and DAG: A response to the comments on our comment. *NeuroImage* **76** 446–449.

LINDQUIST, M. A. and SOBEL, M. E. (2016). Effective connectivity and causal inference in neuroimaging. In *Handbook of Neuroimaging Data Analysis* 419–440. CRC Press, Boca Raton.

LINDQUIST, M. A. and WAGER, T. D. (2007). Validity and power in hemodynamic response modeling: A comparison study and a new approach. *Hum. Brain Mapp.* **28** 764–784.

LINDQUIST, M. A., LOH, J. M., ATLAS, L. Y. and WAGER, T. D. (2009). Modeling the hemodynamic response function in fMRI: Efficiency, bias and mis-modeling. *NeuroImage* **45** S187–S198.

MCLNTOSH, A. and GONZALEZ-LIMA, F. (1994). Structural equation modeling and its application to network analysis in functional brain imaging. *Hum. Brain Mapp.* **2** 2–22.

MEJIA, A., YUE, Y. R., BOLIN, D., LINDREN, F. and LINDQUIST, M. A. (2017). A Bayesian general linear modeling approach to cortical surface fMRI data analysis. Preprint. Available at arXiv:1706.00959.

MIEZIN, F. M., MACCOTTA, L., OLLINGER, J., PETERSEN, S. and BUCKNER, R. (2000). Characterizing the hemodynamic response: Effects of presentation rate, sampling procedure, and the possibility of ordering brain activity based on relative timing. *NeuroImage* **11** 735–759.

MONTI, M. M. (2011). Statistical analysis of fMRI time-series: A critical review of the GLM approach. *Front. Human Neurosci.* **5** 28. https://doi.org/10.3389/fnhum.2011.00028

MOULTON, E. A., SCHMAHMANN, J. D., BECERRA, L. and BORSOOK, D. (2010). The cerebellum and pain: Passive integrator or active participator? *Brains Res. Rev.* **65** 14–27. https://doi.org/10.1016/j.brainresrev.2010.05.005

OGAWA, S., LEE, T.-M., KAY, A. R. and TANK, D. W. (1990). Brain magnetic resonance imaging with contrast dependent on blood oxygenation. *Proc. Natl. Acad. Sci. USA* **87** 9868–9872.

PENNY, W. D., TRUJILLO-BARRETO, N. J. and FRISTON, K. J. (2005). Bayesian fMRI time series analysis with spatial priors. *NeuroImage* **24** 350–362.

POLDRACK, R. A. (2007). Region of interest analysis for fMRI. *Soc. Cogn. Affect. Neurosci.* **2** 67–70. https://doi.org/10.1093/scan/nsm006

POLDRACK, R. A., MUMFORD, J. A. and NICHOLS, T. E. (2011). *Handbook of Functional MRI Data Analysis*. Cambridge Univ. Press, Cambridge. MR2839490 https://doi.org/10.1017/CBO9780511895029

ROBINS, J. M. and HERNÁN, M. A. (2009). Estimation of the causal effects of time-varying exposures. In *Longitudinal Data Analysis. Chapman & Hall/CRC Handb. Mod. Stat. Methods* 553–599. CRC Press, Boca Raton, FL. MR1500133

ROEBROECK, A., FORMISANO, E. and GOEBEL, R. (2005). Mapping directed influence over the brain using Granger causality and fMRI. *NeuroImage* **25** 230–242.

SANYAL, N. and FERREIRA, M. A. (2012). Bayesian hierarchical multi-subject multiscale analysis of functional MRI data. *NeuroImage* **63** 1519–1531.

SOBEL, M. E. (2006). What do randomized studies of housing mobility demonstrate?: Causal inference in the face of interference. *J. Amer. Statist. Assoc.* **101** 1398–1407. MR2307573 https://doi.org/10.1198/016214506000000636

SOBEL, M. E. and LINDQUIST, M. A. (2014). Causal inference for fMRI time series data with systematic errors of measurement in a balanced on/off study of social evaluative threat. *J. Amer. Statist. Assoc.* **109** 967–976. MR3265669 https://doi.org/10.1080/01621459.2014.922886

WAGER, T. D., VAZQUEZ, A., HERNANDEZ, L. and NOLL, D. C. (2005). Accounting for nonlinear BOLD effects in fMRI: Parameter estimates and a model for prediction in rapid event-related studies. *NeuroImage* **25** 206–218.

WAGER, T. D., ATLAS, L. Y., LINDQUIST, M. A., ROY, M., WOO, C.-W. and KROSS, E. (2013). An fMRI-based neurologic signature of physical pain. *N. Engl. J. Med.* **368** 1388–1397.

WOO, C.-W., KRISHNAN, A. and WAGER, T. D. (2014). Cluster-extent based thresholding in fMRI analyses: Pitfalls and recommendations. *NeuroImage* **91** 412–419.

WOO, C.-W., ROY, M., BUHLE, J. T. and WAGER, T. D. (2015). Distinct brain systems mediate the effects of nociceptive input and self-regulation on pain. *PLoS Biol.* **13** e1002036. https://doi.org/10.1371/journal.pbio.1002036

WOOLRICH, M. W., JENKINSON, M., BRADY, J. M. and SMITH, S. M. (2004). Fully Bayesian spatio-temporal modeling of fMRI data. *IEEE Trans. Med. Imag.* **23** 213–231.

YEO, B., KRIENEN, F. M., SEPULCRE, J., SABUNCU, M. R., LASHKARI, D., HOLLINSHEAD, M., ROFFMAN, J. L., SMOLLER, J. W., ZÖLLEI, L. et al. (2011). The organization of the human cerebral cortex estimated by intrinsic functional connectivity. *J. Neurophysiol.* **106** 1125–1165.

ZHANG, L., GUINDANI, M., VERSACE, F., ENGELMANN, J. M. and VANNUCCI, M. (2016). A spatiotemporal nonparametric Bayesian model of multi-subject fMRI data. *Ann. Appl. Stat.* **10** 638–666. MR3528355 https://doi.org/10.1214/16-AOAS926