Functional Causal Mediation Analysis with an Application to Brain Connectivity

Martin A. Lindquist[1]

Department of Statistics, Columbia University, New York, NY 10027

[1]Martin A. Lindquist is Associate Professor, Department of Statistics, Columbia University, New York, NY 10027 (email: *martin@stat.columbia.edu*);

Abstract

Mediation analysis is often used in the behavioral sciences to investigate the role of intermediate variables that lie on the causal path between a randomized treatment and an outcome variable. Typically, mediation is assessed using structural equation models (SEMs) with model coefficients interpreted as causal effects. In this paper we present an extension of SEMs to the functional data analysis (FDA) setting that allows the mediating variable to be a continuous function rather than a single scalar measure, thus providing the opportunity to study the functional effects of the mediator on the outcome. We provide sufficient conditions for identifying the average causal effects of the functional mediators using the extended SEM, as well as weaker conditions under which an instrumental variable estimand may be interpreted as an effect. The method is applied to data from a functional magnetic resonance imaging (fMRI) study of thermal pain that sought to determine whether activation in certain brain regions mediated the effect of applied temperature on self-reported pain. Our approach provides valuable information about the timing of the mediating effect that is not readily available when using the standard non-functional approach. To the best of our knowledge this work provides the first application of causal inference to the FDA framework.

Key words: mediation, structural equation models, functional data analysis, causal inference, fMRI, instrumental variable, brain connectivity

1 Introduction

To date, human brain mapping has been used to primarily construct maps indicating regions of the brain that are activated by certain tasks. Recently, there has been an increased interest in augmenting this type of analysis with connectivity studies that describe how different brain regions interact and how these interactions depend on experimental conditions and behavioral measures. Pathways are considered fundamental properties of brain organization and the ability to understand them is critically important for determining how psychological processes map onto brain function. They provide a means for studying the mechanisms by which experimental manipulations, brain activity, and psychological/physiological outcomes affect one another. In this work we discuss a set of tools for modeling functional pathways using functional magnetic resonance imaging (fMRI) time series data.

Functional MRI measures changes in blood flow and oxygenation in the brain in response to neural stimuli, thereby providing a means to non-invasively study changes in mental activity in response to a certain task (Lindquist, 2008). The technique offers the potential to measure brain activity while experimentally manipulating treatments, thereby providing information beyond the psychological/physiological outcomes of the treatments typically obtained from psychological experiments. A recent example involves the use of brain imaging to study the relationship between a painful thermal stimulus and self-reported pain (Wager et al., 2008). In this experiment a noxious heat stimuli was applied at one of two different levels (high and low) to each of 20 subjects. In response, subjects gave subjective pain ratings at a specific time point following the offset of the stimulus. While the experiment was being performed brain

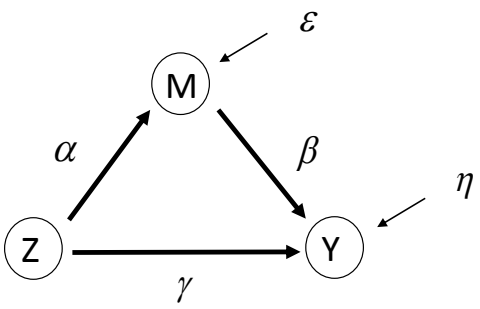

Figure 1: The three-variable path diagram used to represent the standard mediation framework. The variables corresponding to Z, Y and M are all scalar, as are the path coefficients α, β and γ linking them.

activity was measured using fMRI. The goal of the study was to find brain regions whose activity acted as potential mediators of the relationship between temperature and pain rating.

When the effect of the treatment variable Z on the outcome variable Y is at least partially directed through the intervening variable M, then M is said to be a mediator. The three-variable path diagram shown in Fig. 1 is often used to represent such relationships. The influence of the intermediate variable on the outcome is then frequently ascertained using structural equation models, with the model coefficients interpreted as effects. Though the idea of mediation was originally developed in the psychometric and behavioral sciences literature (e.g., Baron and Kenny 1986; MacKinnon 2008), the topic has also received attention in the statistics literature (e.g., Holland, 1988; Robins and Greenland, 1992; Angrist et al., 1996; Ten Have et al., 2007; Albert, 2008; Jo, 2008; Sobel, 2008; VanderWeele, 2009; Imai et al. 2010). The brain imaging experiment can be placed into the three-variable path model by letting the variable Z represent the applied pain level, the variable Y the reported

pain and the variable M the brain response. Here both Z and Y are univariate, while the brain response consists of time series data. In this setting, standard mediation techniques are only applicable if the mediating time course is summarized as a single univariate response, such as peak amplitude or area under the curve. However, the brain response is not necessarily well described by a single summary measure and, in addition, such a measure provides no temporal information about the relationship between $Z - Y$. Therefore, an important extension of the mediation framework would be to allow the variable M to use information across the entire response; something currently not possible in standard mediation models.

In this work we consider the same simple three-variable path model described above, with the novel feature that the intermediate variable M is treated as a continuous function (see Fig. 2 for the analogous path diagram). This provides an extension of current mediation analysis techniques to the functional data analysis (FDA) setting (Ramsay and Silverman, 2005). Conceptually, functional data are thought of as sample paths of a continuous time stochastic process. Although the observed trajectories are often rough and fluctuating, in many applications of FDA there is scientific reason to believe that the true underlying trajectory is a smooth function observed with random error. Extending the mediation framework to the FDA setting allows for the decomposition of the effects of the mediating variable across the support of the function, providing an opportunity to study *functional mediation*.

To assess functional mediation we propose a linear functional structural equation model (lfSEM), which can be used to both test whether an intermediate variable mediates the relationship between treatment and outcome variables and to provide information about the timing of these relationships. We applied the approach to data

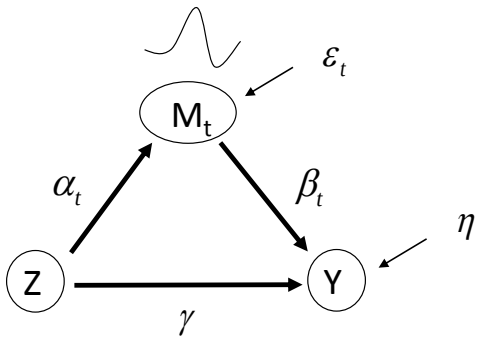

Figure 2: The three-variable path diagram used to represent the functional mediation framework. The variables corresponding to Z and Y are scalars, while the variable corresponding to M is a function. Both the α_t and β_t pathways are represented by functions, while the γ pathway is a scalar.

from the thermal pain experiment and found that we could discriminate between brain regions that mediate the relationship between temperature and pain at the time of the offset of the stimulus and at the time immediately preceding the later pain report. The identification of these separate mechanisms, not possible using standard SEMs, is important as it promises to significantly increase our knowledge about the underlying brain networks associated with different stages of pain processing.

In statistics, randomized experiments are considered the gold standard for estimating causal effects. A common criticism of structural equation models and mediation analysis is that only the assignment to treatment group (i.e. Z) is randomized, while the mediators are self-selected treatments. For this reason the effect of the mediator (the coefficient for the mediator in the regression of the outcome on the mediator and treatment assignment) does not typically warrant a causal interpretation (Holland, 1988). In this work, we adopt the potential outcomes framework for causal inference (Rubin, 1974) to show that under certain assumptions it is possible to obtain

a valid estimate of the average causal effects of the mediator from the parameters of the lfSEM. While a number of researchers have considered different strategies for the identification of direct and indirect effects between treatment and outcome (e.g., Robins and Greenland, 1992; Pearl, 2001; Rubin, 2004; VanderWeele, 2009), we are primarily concerned with determining sufficient conditions required to equate the parameters of lfSEMs with the effects of mediators. A literature has recently developed that deals with identification of direct and indirect effects for situations when interactions and nonlinearities are present (e.g., Peterson et al. 2006; VanderWeele 2009). However, we leave these issues for later work and focus solely on the type of additive models commonly used in the social sciences. We further extend results from Sobel (2008), who deals with the standard non-functional case described in Fig. 1, and provide conditions under which the instrumental variable (IV) estimand may be interpreted as an effect when the intermediate variable is a function.

This article is organized as follows. In Section 2 we use potential outcomes notation to construct a causal linear functional model (CLFM) that allows us to specify the causal parameters of interest used in mediation analysis. These are causal analogues of the parameters shown in the path diagram in Fig. 2. In general, because we are not able to observe all potential outcomes, we cannot directly estimate the parameters of the CLFM. We therefore proceed by discussing a linear functional structural equation model whose parameters can be estimated from the observable data. Under certain assumptions, the parameters of the lfSEM are equivalent to those of the CLFM, thus allowing for the identification of the causal parameters of interest. In Section 3 we compare the IV estimand for the effect of the mediator on the outcome with the equivalent effect in the CLFM. This provides alternative assumptions for

allowing a causal interpretation of the effect of the mediator on the outcome. In Section 4 we describe a technique for estimating the parameters of the lfSEM and the IV-estimand, as well as, simple procedures for performing inference on the model coefficients using resampling methods. Finally, in Sections 5 and 6 we illustrate the utility of the method in a series of simulations and an application to the fMRI study of thermal pain.

2 Identification of Mediated and Unmediated Effects

Consider the path diagram in Fig. 2 where the mediating variable M_t is assumed to be a continuous function. Here the subscript t indicates that a variable/coefficient is a function of t, and we write $\mathbf{M} = \{M_t | t \in [0, 1]\}$ to represent the value of the function over its entire range. In the brain imaging setting, Z is the treatment assignment, Y is reported pain, and M_t is a time series of brain data following each stimulation (that will be treated as samples from a continuous underlying function). Further, the path coefficients α_t and β_t are functions that describe the time-varying relationship between the variables.

This section defines the causal parameters of interest for studying functional mediation, as well as a linear functional SEM that can be used to identify these parameters under conditions to be described. Our goal is to make explicit the assumptions required to equate the parameters of the lfSEM with the causal effects of interest, thus allowing us to determine when the estimates of the former can be given a causal

interpretation. In Section 2.1 we define mediated and unmediated effects at both the subject and population level. A causal model is constructed using potential outcomes for comparison with the lfSEM introduced in Section 2.2. Unlike the parameters of the causal model, the parameters of the lfSEM are identifiable from the observed data. In general, the causal and lfSEM parameters are not equal, implying that the latter should not be interpreted as causal effects. However, we show in Section 2.3 that under certain conditions, the causal parameters of interest are equal to the corresponding parameters of the lfSEM and thus identifiable.

2.1 Defining the Causal Parameters of Interest

In this section we construct a causal model, analogous to the path diagram shown in Fig. 2, using potential outcomes. Consider a randomized experiment consisting of n subjects receiving either the treatment or control condition, where $Z_i = 1$ if subject i is assigned to treatment and 0 otherwise. For each subject i in the population, let $M_{it}(0)$ denote the value of the mediator at time t in the absence of treatment and $M_{it}(1)$ the value at time t under treatment. Similarly, let $Y_i(0, \mathbf{M}_i(0))$ denote i's value on the outcome when assigned to the control group and $Y_i(1, \mathbf{M}_i(1))$ the value if assigned to the treatment group. Throughout, we make the stable unit treatment value assumption (SUTVA; Rubin, 1980) that there is no interference between subjects. This implies that potential outcomes for each subject are unrelated to the treatment given to other subjects, which is reasonable in the brain imaging setting discussed in this paper.

Under SUTVA, the unit-level causal effects of Z on $\mathbf{M}$ and Y, respectively, for

subject i are given by the expressions

$$M_{it}(1) - M_{it}(0) \qquad \forall\, t \tag{1}$$

$$Y_i(1, \mathbf{M}_i(1)) - Y_i(0, \mathbf{M}_i(0)). \tag{2}$$

The latter term can be decomposed as follows:

$$\begin{aligned} Y_i(1, \mathbf{M}_i(1)) - Y_i(0, \mathbf{M}_i(0)) &= \{Y_i(1, \mathbf{M}_i(0)) - Y_i(0, \mathbf{M}_i(0))\} + \{Y_i(1, \mathbf{M}_i(1)) - Y_i(1, \mathbf{M}_i(0))\} \\ &= \{Y_i(1, \mathbf{M}_i(1)) - Y_i(0, \mathbf{M}_i(1))\} + \{Y_i(0, \mathbf{M}_i(1)) - Y_i(0, \mathbf{M}_i(0))\}. \end{aligned} \tag{3}$$

In both decompositions, the first term represents an unmediated (or direct) effect of Z on Y, while the second represents a mediated (or indirect) effect. The direct effects of Z on Y represent the difference in the outcome if one were to change the treatment of subject i from $Z_i = 0$ to $Z_i = 1$, while holding the value of the mediator fixed at $\mathbf{M}_i(0)$ (or $\mathbf{M}_i(1)$). In contrast, the indirect effects represent the difference in the outcomes if the value of the observed mediator were changed from $\mathbf{M}_i(0)$ to $\mathbf{M}_i(1)$ while keeping the actual treatment fixed at 0 (or 1). If the treatment has no effect on the mediator, then $\mathbf{M}_i(1) = \mathbf{M}_i(0)$ and the indirect effects would be zero.

In practice, since only one of the potential outcomes is observable it is not possible to identify unit-level effects. Instead, we average over subjects and seek to estimate population-level effects such as the average causal effect of Z on Y:

$$\tau^{(c)} \equiv E(Y(1, \mathbf{M}(1)) - Y(0, \mathbf{M}(0))). \tag{4}$$

To define other causal effects of interest, we construct a causal linear functional model (CLFM) in which the relationship between the outcome and the mediator follows a

functional linear model (e.g., Ramsay and Silverman, 2005). Potential outcomes are used to express the causal relationship between Z, $\mathbf{M}$ and Y using the system of equations

$$M_{it}(z) = \delta_{1t}^{(c)} + \alpha_t^{(c)} z + \epsilon_{it}(z) \quad (5)$$

$$Y_i(z, \mathbf{m}) = \delta_2^{(c)} + \gamma^{(c)} z + \int_0^1 \beta_t^{(c)} m_t dt + \eta_i(z, \mathbf{m}), \quad (6)$$

where $E(\epsilon_t(z)) = 0 \ \forall \ t$ and $z = 0, 1$ and $E(\eta(z, \mathbf{m})) = 0$ for all values of the pair $(z, \mathbf{m})$. Here the superscript (c) is used to identify the parameters as being causal, in contrast to the analogous parameters $\alpha_t^{(s)}$, $\beta_t^{(s)}$ and $\gamma^{(s)}$ in the linear functional structural equation model presented below.

Using this model we can write the average direct effect of Z on Y, sometimes referred to as the controlled direct effect (VanderWeele, 2009), as

$$E(Y(1, \mathbf{m}) - Y(0, \mathbf{m})) = \gamma^{(c)} \quad (7)$$

for all values of $\mathbf{m}$. This effect differs from the pure direct effect (Robins and Greenland, 1992) which fixes the intermediate variable for each individual to the level it would have received under the absence of treatment. For the CLFM these effects are equivalent under the additional condition $E(\eta(1, \mathbf{M}(0)) - \eta(0, \mathbf{M}(0))) = 0$.

Averaging (3) over all subjects, we obtain

$$\begin{aligned} \tau^{(c)} &= \gamma^{(c)} + E(Y(1, \mathbf{M}(1)) - Y(1, \mathbf{M}(0))) \\ &= \gamma^{(c)} + E(Y(0, \mathbf{M}(1)) - Y(0, \mathbf{M}(0))), \end{aligned} \quad (8)$$

which allows us to express the average total effect as the sum of the average direct and indirect effect of Z on Y. Note that according to (7), the average of both formulations of the direct effect stated in (3) can be expressed using $\gamma^{(c)}$.

Finally, the average causal effect of Z on $\mathbf{M}$, and $\mathbf{M}$ on Y at level $\mathbf{m}$ versus $\mathbf{m}^*$, can be written, respectively, as

$$E(M_t(1) - M_t(0)) \;=\; \alpha_t^{(c)} \qquad \forall\, t \in [0,1] \tag{9}$$

and

$$E(Y(z,\mathbf{m}) - Y(z,\mathbf{m}^*)) \;=\; \int_0^1 \beta_t^{(c)}(m_t - m_t^*)dt \tag{10}$$

Both α_t and β_t depend on t and their value can vary across the range of $[0,1]$, thus allowing the treatment to have a functional effect on the intermediate variable. In our application, this provides a temporal decomposition of the effect that applied pain has on the brain response. This is important as the response can potentially vary with regards to its onset, width or amplitude for different treatments.

The fundamental problem of causal inference (Holland 1988) is that for any given subject one cannot observe the potential outcomes under both the treatment and control conditions at the same time. In addition, terms such as $Y_i(1,\mathbf{M}(0))$ can never be observed regardless of treatment. Hence, randomization of the treatment alone cannot help identify the terms $\gamma^{(c)}$ and $\beta^{(c)}$ and additional assumptions are required to proceed.

2.2 Linear Functional Structural Equation Models

Due to the fundamental problem of causal inference, none of the causal effects defined in the previous section can generally be estimated from the observed data. Therefore, in this section we discuss a linear functional structural equation model, corresponding to the path diagram in Fig. 2, whose parameters can be directly estimated from the data.

A great deal of research has focused on extending linear models to the functional setting (e.g., Hastie and Tibshirani, 1993; Fan and Zhang, 1999; James, 2002; Cardot et al., 2003; Müller and StadtMüller, 2005). Currently, techniques exist for performing regression where both/either the response and explanatory variables are functional rather than scalar. Using these techniques, under conditions described in Section 2.3 mediation can be assessed using a *functional* analogue to a linear structural equation model of the form:

$$M_{it}(Z_i) = \delta_{1t}^{(s)} + \alpha_t^{(s)} Z_i + \epsilon_{it} \qquad \forall t \tag{11}$$

$$Y_i(Z_i, \mathbf{M}_i(Z_i)) = \delta_2^{(s)} + \gamma^{(s)} Z_i + \int_0^1 \beta_t^{(s)} M_{it}(Z_i) dt + \eta_i. \tag{12}$$

where $t \in [0, 1]$. Here the superscript (s) denotes that the parameters are defined in an SEM that is estimable using the observed data. The model parameters are identified through the definitions $E(\epsilon_t | Z = z) = 0 \; \forall t$ and $E(\eta | Z = z, \mathbf{M}(Z) = \mathbf{m}) = 0$. While not necessary for identification, for estimation purposes we also assume that ϵ_{it} and η_i are independent. Using (11) and (12) we can write

$$\alpha_t^{(s)} = E(M_t(1)|Z = 1) - E(M_t(0)|Z = 0) \qquad \forall t \tag{13}$$

$$\gamma^{(s)} = E(Y(Z, \mathbf{M}(Z))|\mathbf{M}(Z) = \mathbf{m}, Z = 1) - E(Y(Z, \mathbf{M}(Z))|\mathbf{M}(Z) = \mathbf{m}, Z = 0) \tag{14}$$

$$\int \beta_t^{(s)}(m_t - m_t^*) dt = E(Y(Z, \mathbf{M}(Z))|\mathbf{M}(Z) = \mathbf{m}, Z = z) - E(Y(Z, \mathbf{M}(Z))|\mathbf{M}(Z) = \mathbf{m}^*, Z = z) \tag{15}$$

In this formulation, it is easy to show that the "total effect" of Z on Y [2]

$$\tau^{(s)} \equiv E(Y(1, \mathbf{M}(1)|Z = 1)) - E(Y(0, \mathbf{M}(0)|Z = 0)) \tag{16}$$

can be expressed as

$$\tau^{(s)} = \gamma^{(s)} + \int_0^1 \alpha_t^{(s)} \beta_t^{(s)} \, dt. \tag{17}$$

Here $\gamma^{(s)}$ represents the "direct effect" of Z on Y, while the term $\int \alpha_t^{(s)} \beta_t^{(s)} \, dt$ represents the "indirect effect". Mediation can then be assessed under suitable conditions by determining whether the integral is significantly different from zero. In addition, the product $\alpha_t^{(s)} \beta_t^{(s)}$ provides a functional decomposition of the "indirect effect" allowing the specific intervals driving the mediation to be determined. Hence, the proposed framework provides the opportunity under suitable conditions to assess the effects of *functional mediation*.

2.3 Comparison of Causal Parameters and lfSEM Parameters

In general, the parameters of the lfSEM ($\tau^{(s)}$, $\gamma^{(s)}$, $\alpha_t^{(s)}$ and $\beta_t^{(s)}$) are not equal to their counterparts in the CLFM ($\tau^{(c)}$, $\gamma^{(c)}$, $\alpha_t^{(c)}$ and $\beta_t^{(c)}$). To equate the parameters we need to make a number of assumptions. We begin by assuming that treatment assignment is ignorable (Rosenbaum and Rubin, 1983).

Assumption 1 *The treatments are assigned independently of the potential outcomes,*

[2] Quotation marks are used in this subsection to differentiate the effects in the lfSEM from their counterparts in the CLFM.

i.e.

$$Y(0, \mathbf{M}(0)), Y(1, \mathbf{M}(1)), \{M_t(0), M_t(1)\}_{t\in[0,1]} \perp\!\!\!\perp Z \tag{18}$$

Under this assumption, it is easy to show that $\tau^{(c)} = \tau^{(s)}$ and $\alpha_t^{(c)} = \alpha_t^{(s)}$. However, in general $\gamma^{(c)} \neq \gamma^{(s)}$ and $\beta^{(c)} \neq \beta^{(s)}$. Under Assumption 1 we can express (15) as

$$\int \beta_t^{(s)}(m_t - m_t^*)\, dt = E(Y(z, \mathbf{m})|\mathbf{M}(Z) = \mathbf{m}) - E(Y(z, \mathbf{m}^*)|\mathbf{M}(Z) = \mathbf{m}^*) \tag{19}$$

which is not directly comparable to the expression in (10). The term $\beta_t^{(s)}$ compares the values of the outcome in two different subpopulations of units, while $\beta_t^{(c)}$ considers the same subjects under different conditions. For similar reasons, the term $\gamma^{(s)}$ is not directly comparable to (7). A sufficient condition for equality of the remaining parameters is to assume that the mediator is ignorable with respect to the potential outcomes.

Assumption 2 *The potential outcomes are independent of the mediator, i.e.*

$$Y(z, \mathbf{m}) \perp\!\!\!\perp \{M_t(z)\}_{t\in[0,1]} \tag{20}$$

for $z = 0, 1$ *and all* $\mathbf{m}$.

Under Assumptions 1-2 it is easy to show the equality of the remaining parameters, as they imply that

$$\begin{aligned} E(Y(Z, \mathbf{M}(Z))|Z = z, \mathbf{M}(Z) = \mathbf{m}) &= E(Y(z, \mathbf{m})|Z = z, \mathbf{M}(z) = \mathbf{m}) \\ &= E(Y(z, \mathbf{m})|\mathbf{M}(z) = \mathbf{m}) \\ &= E(Y(z, \mathbf{m})) \end{aligned} \tag{21}$$

Hence, under Assumptions 1 and 2 $\gamma^{(s)} = \gamma^{(c)}$ and $\beta_t^{(s)} = \beta_t^{(c)}$. Note that these are the same assumptions required for equating the parameters of SEMs with the parameters of an analogous causal model (Sobel, 2008).

In addition, it should also be noted that under Assumption 2 we can write the indirect effect of Z on Y as follows:

$$\begin{aligned} E(Y(0, \mathbf{M}(1)) - Y(0, \mathbf{M}(0))) &= E(Y(1, \mathbf{M}(1)) - Y(1, \mathbf{M}(0))) \\ &= \int_0^1 \alpha_t^{(c)} \beta_t^{(c)} dt. \end{aligned} \tag{22}$$

This, in turn, allows us to express both decompositions of the total effect in (8) as

$$\tau^{(c)} = \gamma^{(c)} + \int_0^1 \alpha_t^{(c)} \beta_t^{(c)} dt \tag{23}$$

providing expressions for the direct and indirect effects in terms of the parameters of the CLFM.

If Assumptions 1 and 2 are valid it is appropriate to make causal claims using the results from the lfSEM. Assumption 1 states that treatment assignment is ignorable, as would be the case in a randomized fMRI experiment. Assumption 2 states that the potential outcomes $Y(z, \mathbf{m})$ are ignorable with respect to the intermediate outcomes $M_t(z)$ given Z, as would be the case if subjects were randomly assigned to $\mathbf{M}$ at both levels of Z. Obviously this is a strong assumption which is untestable in practice, and it is easy to construct examples where it is violated (e.g., Lindquist and Sobel (2011); Lindquist and Sobel (2012)). For example, unobserved variables may exist that confound the relationship between the outcome and the mediator variable even after conditioning on the treatment status, e.g., a latent variable such as pain resilience that simultaneously causes decreased brain response and reported pain. In the next section we look at ways of relaxing Assumption 2.

3 Identification of Causal Effects using an Instrumental Variable

In this section we give conditions under which the instrumental variable (IV) estimand (the effect of Z on Y divided by the effect of Z on $\mathbf{M}$) can be used to identify the causal effect $\beta^{(c)}$. The results follow from similar work by Holland (1988) and Sobel (2008) on SEMs, and provide an alternative set of assumptions for identifying the causal effect of $\mathbf{M}$ on Y. The IV estimand has also been considered in connection with causal effects defined in potential outcomes in work by Angrist and colleagues (Angrist and Imbens, 1995; Angrist et al., 1996).

Under (5) and (6) it is possible to write:

$$\begin{aligned} Y_i(1, \mathbf{M}_i(1)) - Y_i(0, \mathbf{M}_i(0)) &= \gamma^{(c)} + \int_0^1 \beta_t^{(c)}(M_{it}(1) - M_{it}(0))dt \\ &\quad + \eta_i(1, \mathbf{M}_i(1)) - \eta_i(0, \mathbf{M}_i(0)). \end{aligned} \tag{24}$$

To proceed we assume that the difference in potential errors is 0 on average.

Assumption 3 $E(\eta(1, \mathbf{M}(1)) - \eta(0, \mathbf{M}(0))) = 0$.

Because $\tau^{(c)} = \tau^{(s)}$ and $\alpha_t^{(c)} = \alpha_t^{(s)}$ when Assumption 1 holds, using Assumption 3 and averaging over both sides of (24) gives

$$\tau^{(s)} = \gamma^{(c)} + \int_0^1 \alpha_t^{(s)} \beta_t^{(c)} \, dt, \tag{25}$$

which yields one equation in the unknown scalar $\gamma^{(c)}$ and the unknown function $\beta_t^{(c)}$. In some applications it may be reasonable to assume that the treatment effect is transmitted solely through the mediator, leading to the assumption:

Assumption 4 $Y_i(0, \mathbf{m}) = Y_i(1, \mathbf{m}) \qquad \forall\, i,\ \mathbf{m}.$

This assumption, which implies $\gamma^{(c)} = 0$, and is often referred to as an exclusion restriction, states that treatment assignment is unrelated to the potential outcomes once the mediator has been taken into account. Under this assumption we have that

$$\tau^{(s)} \quad = \quad \int_0^1 \alpha_t^{(s)} \beta_t^{(c)}\, dt \tag{26}$$

which can be viewed as an ill-posed homogeneous integral equation of the first kind. In this case the standard IV estimand $\tau^{(s)}/\alpha_t^{(s)}$ is a solution for $\beta_t^{(c)}$, albeit a non-unique one. This particular solution becomes unique if one assumes that the indirect effect is constant across the range of the mediating variable. As the assumption of a constant mediation effect may not be reasonable in fMRI studies, we instead seek a different solution, namely the least-squares solution of minimum norm. We begin by expressing (26) in the form of an operator equation

$$\mathcal{A}\beta^{(c)} \quad = \quad \tau^{(s)} \tag{27}$$

where $\mathcal{A}$ is a linear operator such that $\mathcal{A}\beta^{(c)} = \int \alpha_t^{(s)} \beta_t^{(c)} dt$. This, in turn, can be replaced by the approximating matrix problem

$$\mathbf{A}\tilde{\beta}^{(c)} \quad = \quad \tau^{(s)}, \tag{28}$$

whose solution is given by

$$\tilde{\beta}^{(c)} \quad = \quad \mathbf{A}^{+}\tau^{(s)} \tag{29}$$

where $\mathbf{A}^{+}$ is the Moore-Penrose pseudo-inverse of $\mathbf{A}$ and $\tilde{\beta}^{(c)}$ is the discretized version of $\beta_t^{(c)}$. This solution has many attractive features, including that if $\mathbf{A}$ has real

entries then so does its pseudo-inverse $\mathbf{A}^+$, and if $\mathbf{A}$ is invertible, then $\mathbf{A}^+$ and $\mathbf{A}^{-1}$ coincide. The choice of the minimum norm solution constitutes an additional assumption. However, we differentiate it from the other assumptions due to our belief that they must be verified each time the methodology is applied to a new data set, while the same is not true for the use of the minimum norm solution.

In sum, under Assumptions 1 and 2 we can equate the parameters of the CLFM with the equivalent parameters of the lfSEM. In contrast, under Assumptions 1, 3 and 4 we can use the IV estimand to identify the causal effect of $\mathbf{M}$ on Y. It should be noted that Assumption 2, which states that there is no selection on the mediators with respect to the potential outcomes $Y(z, \mathbf{m})$, is more restrictive than Assumption 3, which states that there is no selection on the difference between potential errors. To see this note that Assumption 2 is equivalent to

$$\eta(z, \mathbf{m}) \perp\!\!\!\perp \{M_t(z)\}_{t\in[0,1]} \tag{30}$$

for $z = 0, 1$ and all $\mathbf{m}$. Hence,

$$\begin{aligned} E(\eta(z, \mathbf{M}(z))) &= E(E(\eta(z, \mathbf{M}(z)))|\mathbf{M}(z) = \mathbf{m}) \\ &= E(E(\eta(z, \mathbf{m}))|\mathbf{M}(z) = \mathbf{m}) \\ &= 0. \end{aligned} \tag{31}$$

which implies Assumption 3. However, Assumption 3 does not imply Assumption 2.

Assumption 4 (exclusion restriction) is not always plausible and the IV estimand will be biased by a factor $\mathbf{A}^+\gamma^{(c)}$ when it does not hold. In this situation assumptions on the sign and/or magnitude of $\gamma^{(c)}$ might be used to bound the causal effect $\beta_t^{(c)}$.

4 Estimation and Inference

In this section we describe an approach for estimating the parameters of the linear functional structural equation model defined in (11)-(12) and the IV estimand defined in (29). We also discuss inferential procedures based on the use of resampling techniques.

Estimation

We begin by describing techniques for estimating the lfSEM. The equation described in (11) consists of a scalar predictor and functional response, while (12) consists of a functional predictor and a scalar response. For both types of models there are a variety of estimation techniques; see Ramsay and Silverman (2006) for a comprehensive overview. In the following we will for simplicity suppress the intercept terms when illustrating the estimation procedure and instead consider the models:

$$M_{it} = \alpha_t^{(s)} Z_i + \epsilon_{it} \tag{32}$$

$$Y_i = \gamma^{(s)} Z_i + \int \beta_t^{(s)} M_{it} dt + \eta_i \tag{33}$$

We approach (32) by expressing $\alpha_t^{(s)}$ as a linear combination of K known basis functions ϕ_{kt}, i.e.

$$\alpha_t^{(s)} = \sum_{k=1}^{K} \phi_{kt} a_k = \Phi_t \mathbf{a} \tag{34}$$

where $\Phi_t = (\phi_{1t}, \cdots \phi_{Kt})$ and $\mathbf{a} = (a_1, \cdots a_K)^T$. The coefficients of the expansion, $\mathbf{a}$, are determined by minimizing the least squares criterion

$$LMSSE(\alpha_t^{(s)}) = \sum \int [M_{it} - \alpha_t^{(s)} Z_i]^2 dt. \tag{35}$$

If we re-express $M_{it} - \alpha_t^{(s)} Z_i$ as $M_{it} - Z_i \Phi_t \mathbf{a}$ and let $\Psi_{it} = Z_i \Phi_t$, then the solution is given by

$$\hat{\mathbf{a}} = \left[\sum \int \Psi_{it}^T \Psi_{it} dt \right]^{-1} \left[\sum \int \Psi_{it}^T M_{it} dt \right], \tag{36}$$

which provides us with an estimate of $\alpha_t^{(s)}$.

The space of functions $\beta_t^{(s)}$ satisfying (33) is infinite-dimensional and simply minimizing the sum of squares will not provide a meaningful estimate. Instead, we minimize the penalized square error

$$PENSSE_\lambda(\gamma^{(s)}, \beta_t^{(s)}) = \sum \left[Y_i - \gamma^{(s)} Z_i - \int \beta_t^{(s)} M_{it} dt \right]^2 + \lambda \int D^2 \beta_t^{(s)} dt, \tag{37}$$

where the integral portion of the second term represents the integrated squared second derivative and λ represents a constant smoothing parameter. We proceed by expressing the regression function as

$$\beta_t^{(s)} = \sum_{k=1}^{K} \phi_{kt} b_k = \Phi_t \mathbf{b} \tag{38}$$

where $\mathbf{b} = (b_1, \cdots b_K)^T$. Using this expression we can write (33) as

$$\begin{aligned} Y_i &= \gamma^{(s)} Z_i + \int \beta_t^{(s)} M_{it} dt + \eta_i \\ &= \gamma^{(s)} Z_i + \left[\int \Phi_t M_{it} dt \right] \mathbf{b} + \eta_i \\ &= \gamma^{(s)} Z_i + \Lambda_i \mathbf{b} + \eta_i. \end{aligned} \tag{39}$$

where $\Lambda_i = \int \Phi_t M_{it} dt$. Concatenating the results across all subjects we obtain

$$Y = G\xi + \eta \tag{40}$$

where

$$Y = \begin{pmatrix} Y_1 \\ \vdots \\ Y_n \end{pmatrix}, G = \begin{pmatrix} Z_1 & \Lambda_1 \\ \vdots & \vdots \\ Z_n & \Lambda_n \end{pmatrix} \text{ and } \xi = \begin{pmatrix} \gamma^{(s)} \\ \mathbf{b} \end{pmatrix}. \tag{41}$$

The solution of (37) is given by

$$\hat{\xi} = (G^T G + \lambda R)^{-1} G^T Y, \tag{42}$$

where R is a matrix with elements R_{ij} equal to zero if j or $k = 1$ and $\int D^2\phi_{js} D^2\phi_{ks} ds$ otherwise.

To compute the IV-estimand, we first obtain the path coefficient $\alpha_t^{(s)}$ as outlined above. The discretized version of the linear operator $\mathcal{A}$, denoted $\mathbf{A}$, can then be obtained by evaluating $\Phi_t \hat{\mathbf{a}}$, $t \in [0, 1]$, at N equidistant points. Next, the path coefficient $\tau^{(s)}$ is computed using the equation

$$Y_i(Z_i, \mathbf{M}_i(Z_i)) \;\; = \;\; \delta_3^{(s)} + \tau^{(s)} Z_i + \epsilon_i, \tag{43}$$

which can be solved using standard linear regression methods. Finally, the IV estimand is obtained by taking the pseudo inverse of $\mathbf{A}$ and computing $\mathbf{A}^+ \tau^{(s)}$.

Inference

Once the parameters of the linear functional SEM and/or the IV-estimand have been estimated, we can perform inference on the indirect effect. In brain imaging, hypothesis testing is considered to be of primary importance. In mediation analysis we are particularly interested in determining whether the total effect is stronger than the direct relationship controlling for $\mathbf{M}$. Analogous to the fully univariate setting (Sobel, 1982; Baron and Kenny, 1986), this can be done by testing the significance of

the integral of the $\alpha_t^{(s)}\beta_t^{(s)}$ product. As the exact distribution of the function $\alpha_t^{(s)}\beta_t^{(s)}$ is unknown, we use bootstrap methods to perform inference, providing a functional equivalent to an approach often used in the univariate setting (Shrout and Bolger, 2002).

We propose using a wild bootstrap procedure as it is known to perform well for regression models (Flachaire, 2003; Zhu et al., 2007). To produce a bootstrap sample $(Z_i,\ Y_i^*,\ \mathbf{M}_i^*)$, $i = 1, \ldots n$, we use the following data generating process (DGP):

$$M_{it}^*(Z_i) = \hat{\delta}_{1t}^{(s)} + \hat{\alpha}_t^{(s)} Z_i + v_i^* \hat{\epsilon}_{it} \qquad \forall t \tag{44}$$

$$Y_i^*(Z_i, \mathbf{M}_i^*(Z_i)) = \hat{\delta}_2^{(s)} + \hat{\gamma}^{(s)} Z_i + \int_0^1 \hat{\beta}_t^{(s)} M_{it}^*(Z_i) dt + v_i^* \hat{\eta}_i. \tag{45}$$

where $\hat{\delta}_{1t}^{(s)}$, $\hat{\delta}_2^{(s)}$, $\hat{\alpha}_t^{(s)}$, $\hat{\beta}_t^{(s)}$ and $\hat{\gamma}^{(s)}$ are estimates of the parameters in (11) and (12), $\hat{\epsilon}_{it}$ and $\hat{\eta}_i$ are the i^{th} residuals and v_i^* are independent and identically distributed as

$$v_i^* = \begin{cases} 1 & \text{with probability } 0.5 \\ -1 & \text{with probability } 0.5 \end{cases}$$

This particular DGP has been studied extensively in the context of IV regression (Davidson and MacKinnon, 2008) and shown to outperform other comparable bootstrap methods.

The bootstrap procedure is performed as follows:

1. Independently generate a bootstrap sample $(Z_i,\ Y_i^*,\ \mathbf{M}_i^*)$, $i = 1, \ldots n$, using the wild bootstrap DGP described in (44) and (45).

2. Using the resampled data, refit (11) and (12) and record the estimate of $\alpha_t^{(s)}\beta_t^{(s)}$ for $t \in [0, 1]$. When working with the IV-estimand, refit (29) rather than (12).

3. Repeat the procedure outlined in steps 1 and 2 B (e.g., $1,000$) times.

The B replications are used to compute the bootstrap distribution of $\alpha_t^{(s)}\beta_t^{(s)}$ for all values of t, which can be used to test whether $\alpha_t^{(s)}\beta_t^{(s)}$ differs significantly from 0. Because the suggested DGP does not impose the null hypothesis on the resampled data, it is necessary to instead test whether $\alpha_t^{(s)}\beta_t^{(s)}$ differs from $\hat{\alpha}_t^{(s)}\hat{\beta}_t^{(s)}$ to obtain a valid test. It should be noted that the bootstrap distribution can also be used to compute percentile confidence intervals (Efron and Tibshirani, 1998) for $\alpha_t^{(s)}\beta_t^{(s)}$ for all t.

Because we are interested in testing for significance across the range of t, it is necessary to correct for multiple comparisons. Throughout the manuscript we use the Benjamini-Hochberg procedure (Benjamini and Hochberg, 1995) to control the false detection rate (FDR). We use FDR contolling methods, rather than methods that control the family-wise error rate (FWER), because choosing the appropriate threshold to control the FWER is difficult due to dependencies in the data.

5 Simulation study

In this section we report the results of a series of simulation studies where data was generated using different combinations of the assumptions described in Sections 2-3. In the first five simulations (summarized in the schematic shown in Fig. 3) data was generated according to the model described in (11) and (12). This is meant to represent a situation under which sequential ignorability (Assumptions 1-2) holds. The first three simulations were designed to determine the empirical false positive rate of the method when no mediation effect is present, while the next two were designed to study the power in the presence of a significant mediation effect. The sixth and

final simulation, generated using the causal model (5) and (6), was designed to mimic a situation where Assumption 2 is violated, but the assumptions for identifying $\beta_t^{(c)}$ using the IV-estimand hold (Assumptions 3-4).

In five of the simulations ($1-4$, 6), data was generated for 20 "subjects", 10 randomly assigned to a treatment group ($Z=1$) and 10 to a control group ($Z=0$). In the fifth simulation, data was generated for n "subjects", with n allowed to vary between 10 and 50, split equally between $Z=0$ and $Z=1$. The values of M_t and Y differed between the six simulations as outlined below and were chosen to mimic plausible brain imaging settings. Throughout, we restricted M_t to a uniform grid of 60 points in the interval $[0,1]$. In five of the simulations the shape of M_t is based on a common time-varying function denoted h_t (see bottom-left hand side of Fig. 3) which equals 0 in the ranges $[0,1/3)$ and $(2/3,T]$, and varies smoothly in the range $[1/3,2/3]$. The function corresponds to the so-called canonical hemodynamic response function (HRF; Lindquist, 2008) typically used to model changes in fMRI signal in response to a stimulus. This function consists of an early rise in activation level, followed by a decline and a subsequent post-activation undershoot below baseline level.

Each of the six simulations was repeated 500 times. For each repetition of Simulations $1-5$ the models (11) and (12) were fit using the approach outlined in Section 4. When fitting the lfSEM model we restrict M_t to a uniform grid of 60 points in the interval $[0,1]$. The path coefficients $\alpha_t^{(s)}$ and $\beta_t^{(s)}$ are modeled using a b-spline basis set of order 6 with 30 knots. The path equations are computed using custom Matlab software, incorporating functions from the Matlab toolbox for functional data analysis[3]. The roughness penalty required to fit (12) is taken as the L_2-norm of its second

[3]http://www.psych.mcgill.ca/misc/fda/index.html

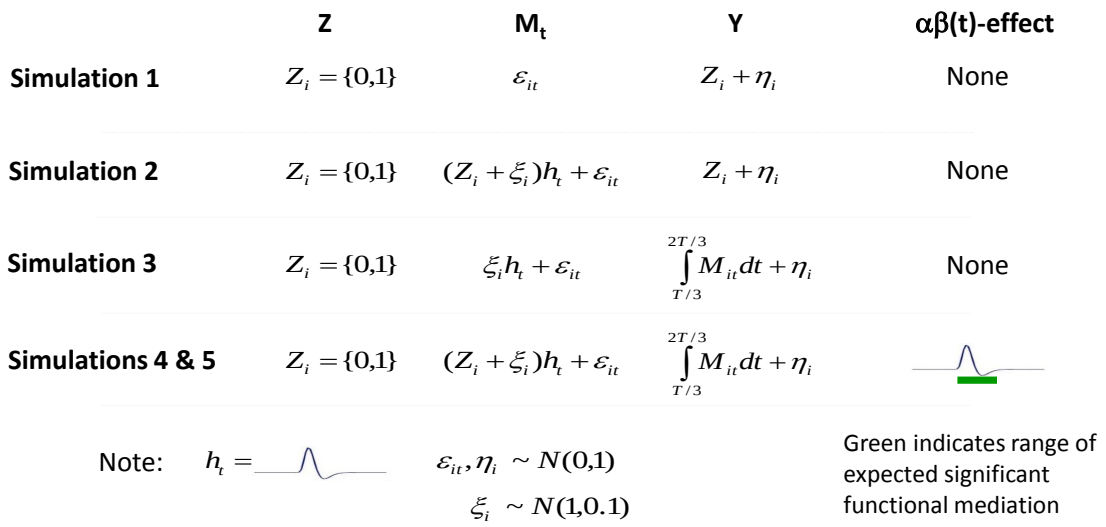

Figure 3: A schematic overview of the first five simulation studies along with expected $\alpha_t^{(s)}\beta_t^{(s)}$ effects (last column). Simulations $1-3$ were designed to determine the empirical false positive rate of the method when no mediation effect is present, while Simulations $4-5$ were designed to study the power in the presence of a mediation effect. In Simulations $1-4$, data was generated for 20 "subjects", where half are assigned to a treatment group ($Z=1$) and half to a control group ($Z=0$). In Simulation 5 the sample size was allowed to vary between 10 and 50 "subjects". The values of M_t and Y differed between the five simulations as summarized in the schematic in order to mimic plausible brain imaging settings.

derivative, with the smoothing parameter selected by leave-one-out cross-validation; no smoothing is used in the initial step of representing M_t in terms of the B-spline basis. A bootstrap test was performed to determine time points where $\alpha_t^{(s)}\beta_t^{(s)}$ was significantly different from 0. The results were controlled for multiple comparisons using the Benjamini-Hochberg procedure ($q=0.05$).

For each repetition of Simulation 6, $\alpha_t^{(c)}\beta_t^{(c)}$ is obtained by equating $\alpha_t^{(c)}=\alpha_t^{(s)}$ and computing the IV estimand $\beta_t^{(c)}$ using (29). A bootstrap test was performed to determine whether $\alpha_t^{(c)}\beta_t^{(c)}$ was significantly different from 0, controlling for multiple comparisons using the Benjamini-Hochberg procedure ($q=0.05$).

Simulation 1*:* The data is generated assuming $M_{it} = \epsilon_{it}$ and $Y_i = Z_i + \eta_i$ for all $i = 1, \ldots 20$, where both ϵ_{it} and η_i follow a standard normal distribution. In this simulation there should be a significant relationship between Z and Y, but it should not be mediated by M_t. This holds because there is no effect of Z on M_t and similarly no effect of M_t on Y, i.e. $\alpha^{(s)} = \beta^{(s)} = 0$. The first row of Fig. 4 shows estimates of $\alpha_t^{(s)}$, $\beta_t^{(s)}$ and $\alpha_t^{(s)}\beta_t^{(s)}$ together with 95% bootstrap percentile confidence intervals. None of the effects appear to deviate significantly from 0 across the time interval. Fig. 5A shows the proportion of times $\alpha_t^{(s)}\beta_t^{(s)}$ was deemed significant as a function of t in the 500 replications. Clearly, all time points fall well below 0.05.

Simulation 2*:* The data is generated in an analogous manner as in Simulation I, except $M_{it} = (Z_i + \xi_i)h_t + \epsilon_{it}$ for all $i = 1, \ldots 20$, where h_t is defined as above, ϵ_{it} follows a standard normal distribution and ξ_i is $N(1, 0.1)$. Again there should be a significant relationship between Z and Y, but it should not be mediated by M_t as there is no effect of M_t on Y since $\beta_t^{(s)} = 0$ for all t. Note that in this case $\alpha_t^{(s)} \neq 0$, differentiating it from the setting of Simulation I. The second row of Fig. 4 shows estimates of $\alpha_t^{(s)}$, $\beta_t^{(s)}$ and $\alpha_t^{(s)}\beta_t^{(s)}$ together with 95% bootstrap percentile confidence intervals. The estimates coincide with the simulated values, with only $\alpha_t^{(s)}$ being significantly different from 0 in the time interval corresponding to where h_t is non-zero. Fig. 5B shows the proportion of times $\alpha_t^{(s)}\beta_t^{(s)}$ was deemed significant as a function of t. Again, all time points fall below 0.05. However, the proportion of false positives are increased and approaches 0.05 in the interval where h_t is non-zero.

Simulation 3*:* The data is generated as $M_{it} = \xi_i h_t + \epsilon_{it}$ and $Y_i = \int_{1/3}^{2/3} M_{it}dt + \eta_i$ for all $i = 1, \ldots 20$, where h_t is defined as above and both ϵ_{it} and η_i follow standard normal

distributions and ξ_i is $N(1, 0.1)$. In this simulation there should be no relationship between Z and Y or Z and M as $\tau^{(s)} = 0$ and $\alpha_t^{(s)} = 0$ for all t. The third row of Fig. 4 shows estimates of $\alpha_t^{(s)}$, $\beta_t^{(s)}$ and $\alpha_t^{(s)}\beta_t^{(s)}$ together with 95% bootstrap percentile confidence intervals. Again, the estimates seem to coincide with the simulated values, with only $\beta_t^{(s)}$ being significantly different from 0 in the time interval $[1/3, 2/3]$. Fig. 5C shows the proportion of times $\alpha_t^{(s)}\beta_t^{(s)}$ was deemed significant as a function of t in the 500 repetitions. Again, all time points fall well below 0.05.

Simulation 4*:* The data is generated in an analogous manner as described in Simulation III, except $M_{it} = (Z_i + \xi_i)h_t + \epsilon_{it}$ for all $i = 1, \ldots 20$, where ϵ_i follows a standard normal distribution and ξ_i is $N(1, 0.1)$. In this simulation the relationship between Z and Y is significant and should be mediated by M_t in the range $[1/3, 2/3]$ as both $\alpha_t^{(s)}$ and $\beta_t^{(s)}$ are non-zero in that interval. The fourth row of Fig. 4 shows estimates of $\alpha_t^{(s)}$, $\beta_t^{(s)}$ and $\alpha_t^{(s)}\beta_t^{(s)}$ together with 95% bootstrap percentile confidence intervals. Again, all three estimates coincide with the simulated values, as they are significantly different from 0 in the appropriate intervals of time. Fig. 6A shows the proportion of times $\alpha_t^{(s)}\beta_t^{(s)}$ was deemed significant as a function of t. In the intervals $[0, 1/3)$ and $(2/3, 1]$ all time points fall below 0.05. In addition, we see a significant relationship in the range $[1/3, 2/3]$ as expected. The proportion significant results peaks at 45% around the point where h_t takes its maximum.

Simulation 5*:* The data is generated in an analogous manner as described in Simulation IV except the sample size was allowed to vary between 10 and 50 "subjects". Here the relationship between Z and Y should be mediated by M_t in the range $[1/3, 2/3]$ as before. Fig. 6B shows the proportion of times $\alpha_t^{(s)}\beta_t^{(s)}$ was deemed significant as

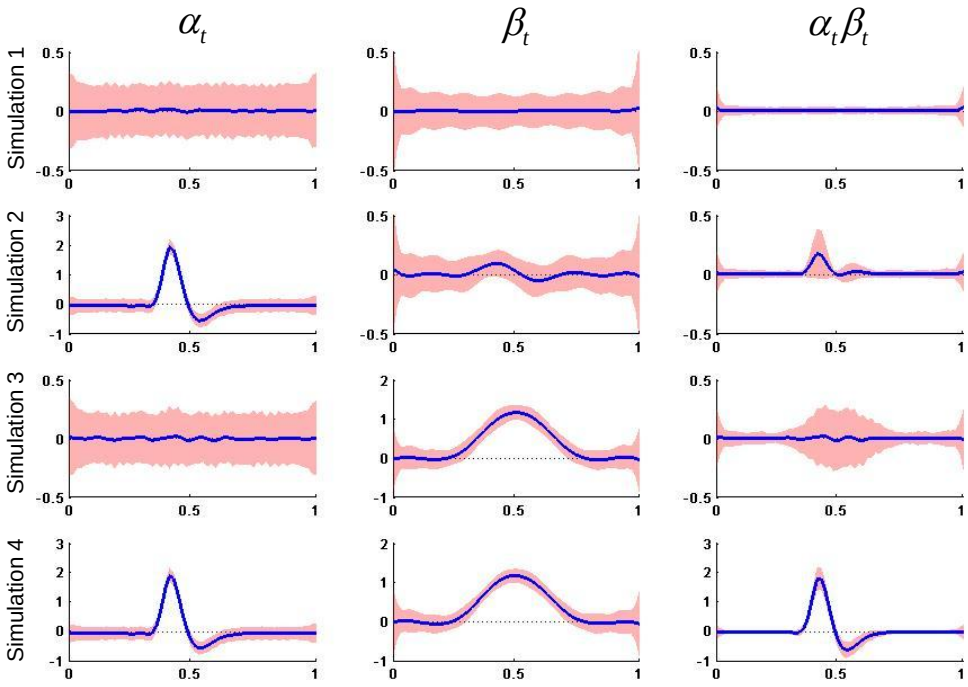

Figure 4: Estimates of $\alpha_t^{(s)}$, $\beta_t^{(s)}$ and $\alpha_t^{(s)}\beta_t^{(s)}$ together with uncorrected 95% bootstrap percentile confidence intervals for data generated according to the settings described in the first four simulations studies.

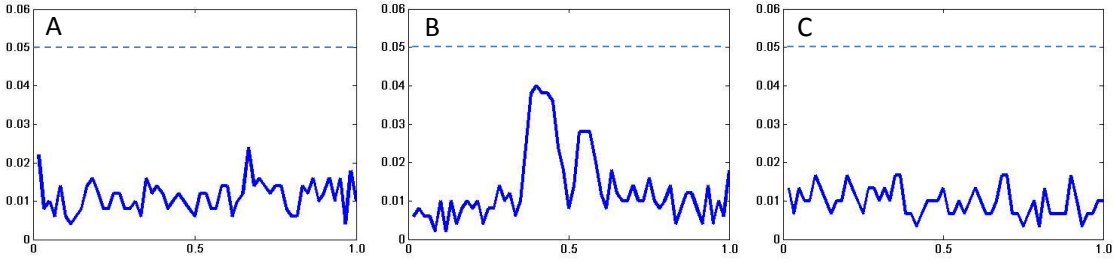

Figure 5: Results of the first three simulation studies are shown in (A)-(C). Each plot shows the proportion of times $\alpha_t^{(s)}\beta_t^{(s)}$ was deemed significant as a function of time. The results illustrate that the method provides adequate control of the false positive rate in all three simulated scenarios.

a function of t for a number of sample sizes ranging from 10 to 50. As expected the higher the sample size, the better the sensitivity and specificity.

Simulation 6*:* The data is generated according to the causal model in (5) and (6) with $\delta_2^{(c)} = \gamma^{(c)} = 0$, $\delta_{1t}^{(c)} = 0\ \forall t$, $\alpha_t^{(c)} = h_t$ and $\beta_t^{(c)} = 1$ in the range $[2/6, 3/6]$ and 0 otherwise. We ensured that $\epsilon_{it}(0), \epsilon_{it}(1) \perp\!\!\!\perp Z_i$ (Assumption 1) by generating both potential outcomes from a standard normal model and randomly assigning the observed outcome to each subject. Further we set $\eta_i(z, \mathbf{m}) = \int_{2/6}^{3/6} \epsilon_{it}(z)dt + \eta_i^*(z, \mathbf{m})$ where $\eta_i^*(z, m)$ is standard normal. This ensures that Assumption 2 is violated since $\eta_i(z, m)$ is not independent of $\epsilon_i(z)$ for $z = 0,\ 1$. However, since Assumptions 3-4 hold we can estimate $\beta_t^{(c)}$ using the IV estimand. In this simulation the relationship between Z and Y is significant and should be mediated by M_t in the range $[2/6, 3/6]$ as both $\alpha_t^{(c)}$ and $\beta_t^{(c)}$ are non-zero in that interval. Fig. 6C shows the proportion of times $\alpha_t^{(c)}\beta_t^{(c)}$ was deemed significant as a function of t. In the intervals $[0, 2/6)$ and $(3/6, 1]$ almost all time points fall below 0.05. In addition, we see a significant relationship in the range $[2/6, 3/6]$ as expected. The proportion significant results almost reaches 100% around the point where h_t takes its maximum.

6 Experimental Data

In this section we study data from the fMRI study of thermal pain ($n = 20$) described in the Introduction. Functional MRI data was extracted from 21 different classic pain-responsive brain regions. Each time course consisted of 23 equidistant temporal measurements made every 2 s, providing a total of 46 s of brain activation ranging from the time of the application of the heat stimuli to the pain report. The stimuli

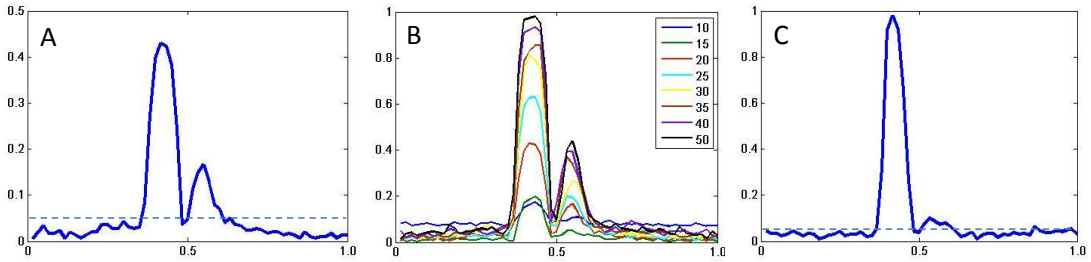

Figure 6: Results of the last three simulation studies. (A) The plot shows the results of the fourth simulation. It shows the proportion of times $\alpha_t^{(s)}\beta_t^{(s)}$ was deemed significant as a function of time and illustrates the power of the method in detecting true positives in the interval $[1/3, 2/3]$ while appropriately controlling for false positives in the intervals $[0, 1/3]$ and $[2/3, 1]$. (B) The plot illustrates the power to detect true positives as a function of sample size, with values ranging from 10 to 50. As expected the higher the sample size, the better the sensitivity and specificity. (C) The plot shows the results of the sixth simulation and illustrates the power of the IV estimand to detect true positives.

consisted of thermal stimulations delivered to the left volar forearm that participants judged to be non-painful vs. near the limit of pain tolerance. The temperature of these painful (i.e., hot) and non-painful (i.e., warm) stimulations was determined via a pain calibration task that took place prior to the experiment on the day of scanning. Following an 18 s interval of thermal stimulation, a fixation cross was presented for a 14 s interval until the words "How painful?" appeared on the screen. After a few seconds of contemplation the participants rated the overall pain intensity on a 10-point numerically anchored visual analog scale (VAS), similar to those commonly used in clinical practice. Participants respond by indicating a position along a continuous line between two end-points. The continuous aspect of the scale differentiates it from similar discrete scales (e.g. the Likert scale), and allows us to use the suggested model instead of a variant specifically designed for ordinal responses.

Each time course was placed into the three-variable path model shown in Fig. 2;

where the variable Z represents the applied pain level, the variable Y the reported pain and the variable M the brain response. We began by estimating the parameters using the lfSEM framework. We restrict M_t to a uniform grid of 23 points in the range $[0, 46]$. The path coefficients $\alpha_t^{(s)}$ and $\beta_t^{(s)}$ were modeled using a b-spline basis set of order 6 with 10 equidistantly spaced knots. Inference was performed using a bootstrap test to determine whether $\alpha_t^{(s)}\beta_t^{(s)}$ was significantly different from 0 in any time interval between pain application and pain report. The Benjamini-Hochberg procedure was used to control the false detection rate ($q = 0.05$). Next, we estimate the IV estimand using the technique outlined in Section 4.

6.1 Results of functional Mediation Analysis

Fig. 7 shows results from the right anterior insula which has been shown to be related to negative emotional experience. This region is believed to mediate the relationship between temperature and pain rating. Estimates of the $\alpha_t^{(s)}$ and $\beta_t^{(s)}$ pathway functions are shown on either side of the path model. These functions suggest a sustained period for which activity is modulated by temperature, and a more phasic response predicting perceived pain, controlling for temperature. A 95% bootstrap percentile confidence interval shows that the $\alpha_t^{(s)}\beta_t^{(s)}$-effect is significantly non-zero in the range between $20 - 24$ seconds after the start of the trial, indicating the key time interval driving the mediation. This corresponds to the first 4 seconds following the end of the application of heat. This somewhat delayed effect is due to the sluggish nature of brain hemodynamics, which peaks roughly 6 seconds after peak neuronal activation, and is consistent with timings of other fMRI experiments

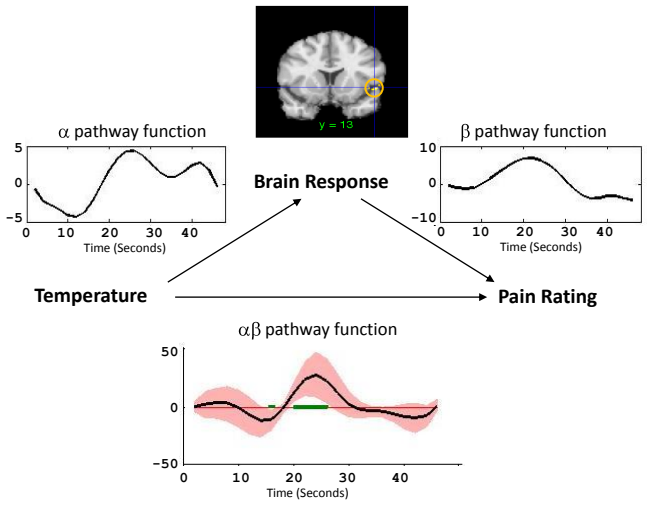

Figure 7: Results show that activation in the right anterior insula (see brain map for the anatomical location of the region) mediates the relationship between temperature and pain rating. Estimates of the $\alpha_t^{(s)}$ and $\beta_t^{(s)}$ pathway functions are shown on either side of the path model. These functions suggest a sustained period for which activity is modulated by temperature, and a more phasic response predicting perceived pain, controlling for temperature. An uncorrected 95% bootstrap percentile confidence interval shows that the $\alpha_t^{(s)}\beta_t^{(s)}$-effect is significantly non-zero in an interval between $20 - 24$ seconds following the start of the trial (shown in red in bottom plot). This indicates the key time interval driving the mediation.

(Lindquist, 2008).

Fig. 8 shows time points with significant non-zero $\alpha_t^{(s)}\beta_t^{(s)}$-effects for each of 21 brain regions obtained using a bootstrap test. The results show that data in many classic "pain-responsive regions" such as the anterior insula (AINS) show significant mediation of the temperature-report relationship particularly around the end of the treatment, i.e. heat application ($20 - 24$ sec). The subjective pain grows during the stimulation period, and is often maximal around the end of stimulation, which typically drives pain ratings made after the trial. Notably, some other regions show significant mediation effects around the time immediately preceding pain reporting

$(38 - 44\ s)$, perhaps signaling a contribution of activity during "pain recall". In particular, regions in the insular cortex appear to be active during pain judgment, which is reasonable due to their link to decision-making (Moulton et al. 2005; Grinband et al. 2006). In addition to responding to physical pain, the anterior insula has been shown to be activated by viewing others in pain, receiving an unfair offer in an economic game, or viewing aversive pictures (Wager et al., 2008; Amodio et al., 2006).

The ability to identify these disjoint time intervals as being involved in mediating the effect between temperature and reported pain is a unique property of the functional mediation framework and is not possible using standard mediation techniques. Hence, we believe that the application of lfSEM provides invaluable information about different components of pain processing in the human brain. This, in turn, furthers our understanding of the components involved in the processing and reporting of applied pain.

Fig. 9 shows the IV estimands for each of the 21 brain regions. The results show that for most regions there are three distinct time intervals of significant effects. The first, roughly $4 - 16\ s$ following the start of the trial, corresponds to a negative value of the IV estimand. Hence, high brain activation in the early portion of the trial has a negative effect on reported pain. This time period is typically thought of as a baseline period before task related activation has had time to appear and any activation may be related to anticipation. In contrast the time periods $20 - 28\ s$ and $38 - 44\ s$ corresponds to a positive value of the IV estimand. In these intervals high activity has a positive effect on reported pain, and they are related to the timing of task-related activation and pain recall, respectively, as discussed above.

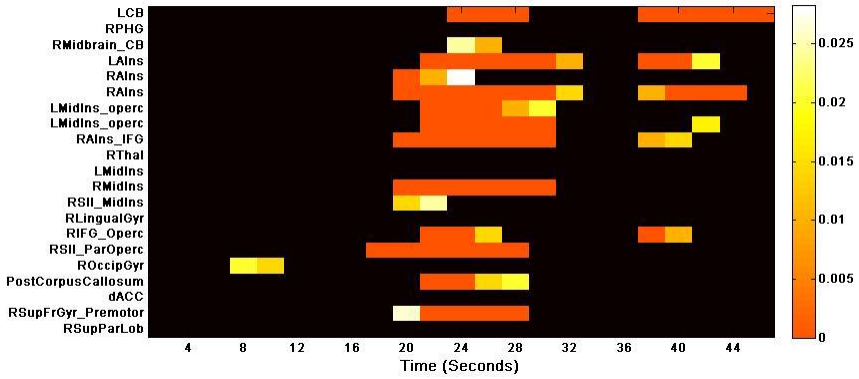

Figure 8: Time points with significant non-zero $\alpha_t^{(s)}\beta_t^{(s)}$-effects for each of 21 brain regions (FDR corrected p-value < 0.05). The colorbar shows the size of the p-value. The results show that classic "pain-responsive regions" show significant mediation of the temperature-report relationship particularly around the end of the treatment $(20 - 24\ s)$. Some other pain regions show significant mediation effects around the time of pain reporting, perhaps signaling a contribution of activity during pain recall.

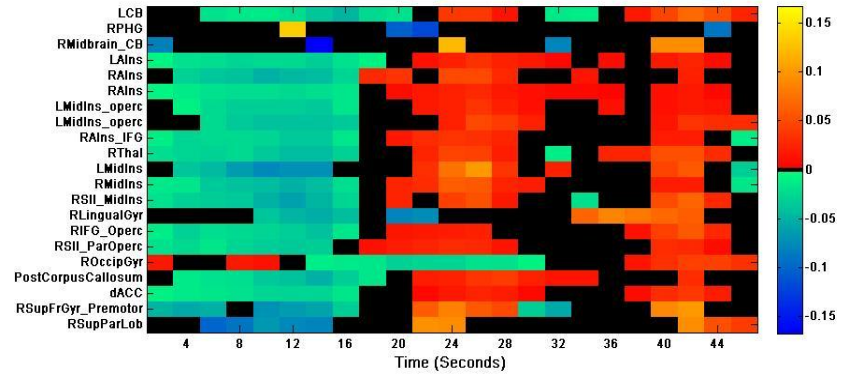

Figure 9: Time points with significant $\beta_t^{(c)}$-effects for each of 21 brain regions (FDR corrected p-value < 0.05). The color bar shows the amplitude of the estimate, with results obtained using the IV estimand.

6.2 Assessment of assumptions

The validity of the causal model defined in Section 2.1 depends upon the appropriateness of SUTVA and models (1) and (2). In our application neither the brain response nor the reported pain of an individual subject should be affected by the treatment given to any other subject in the study, thus validating SUTVA. In addition, linear additive models are commonly used in the analysis of neuroimaging data and their usage in this particular application appears reasonable. To partially verify linearity and additivity we studied a number of independent data sets that used the same paradigm and matched values of the mediator to compute $Y_i(1, \mathbf{m}) - Y_i(0, \mathbf{m})$ for different values of $\mathbf{m}$. The results indicate only minor deviations from linearity or additivity.

To apply a causal interpretation to the parameters of the linear functional structural equation model and the IV estimand we require certain combinations of Assumptions 1-4 to hold. The requirements for equating the parameters of the lfSEM with the effects of the mediators are given by Assumptions 1-2. As treatments are randomly assigned to subjects it is reasonable to assume that the treatment is ignorable. Assumption 2 would be valid if the mediators were randomly assigned to the subjects. However, this is not the case here and instead we must assume that they behave as if they were. This assumption is unverifiable in practice and ultimately depends on context. In the neuroimaging setting its validity may differ across brain regions, making causal claims more difficult to access. For example, assume subjects in the study are either resilient or not to pain. Suppose that resilient subjects tend to have a lower brain response than non-resilient subjects when a painful stimulus

is applied. Further suppose that resilient subjects always report lower pain ratings than their non-resilient counterparts at any value of the mediating variable. In this situation, $Y(1, \mathbf{m})$ will not be independent of $\mathbf{M}(1)$ and Assumption 2 will be violated. In practice, this assumption could potentially be weakened by allowing for conditioning on potential confounders. However, no such covariates were available in this particular study.

For a valid causal interpretation of the IV estimand, we require Assumptions 1, 3 and 4. The assumption of no average difference in potential errors appears plausible, in particular in comparison with Assumption 2. Revisiting the example with resilient/non-resilient subjects, here it suffices that the resilient (non-resilient) subjects on average fall above (below) the mean pain rating by the same amount under both conditions. Finally, Assumption 4 is untestable in practice but amounts to attributing the effect of Z on Y to the change in the brain response rather than the change in the treatment. One caveat is that since a single mediator model is assumed and multiple mediators have been targeted in the study, this assumption may potentially be violated as the effects of omitted mediators may become confounded with the direct effect. Ideally, we would wish to use a model that simultaneously considers the effects of multiple mediators on the outcome to circumvent this issue. We leave this for future research and simply note that any violation of Assumption 4 gives rise to bias. As a final note on assumptions, the application considers M_{it} (and the corresponding coefficients) as functions of time. Hence, it is conceivable that values of the mediator at earlier times could affect values at later times. The implicit assumption of our model is that there is no causal connection among the multiple measurements of the mediator, i.e. that they behave as if they were measured

contemporaneously.

7 Discussion

This article introduces linear functional structural equation modeling as a means of studying time-varying mediation effects. It also examines conditions under which the lfSEM model and instrumental variable methods can be used to identify causal effects of a mediating functional variable on an outcome. To the best of our knowledge this work provides the first application of causal inference to the FDA framework.

The causal interpretation for the parameters of the lfSEM rests on a strong untestable assumption, namely sequential ignorability (Assumptions 1-2). Recent advances in imaging may allow us to avoid Assumption 2 as neuroimaging data has been combined with transcranial magnetic stimulation (TMS) to integrate the ability of neuroimaging to observe brain activity with the ability of TMS to manipulate brain function (Bohning et al., 1997). Using this technique one can simulate temporary "brain lesions" while the subject performs certain tasks. One can then attempt to infer causal relationships by studying differences in a brain network when a region is functioning and when it is not. Using this technique will allow us to experimentally manipulate the mediating variable, as well as the treatment, allowing us to circumvent the sequential ignorability assumption.

The proposed lfSEM framework assumes that the treatment and outcome variables are both univariate, while the intermediate variable is continuous. The method can be extended to allow the variables Z, $\mathbf{M}$ and Y to all be functional. There are several ways to formulate an analogous system of equations to (11) and (12) for this case. One

example which uses the time-varying, or concurrent, model for functional regression (Hastie and Tibshirani, 1993; Ramsey, Hooker and Graves, 2009) can be expressed as follows:

$$M_{it}(Z_{it}) = \delta_{1t} + \alpha_t Z_{it} + \epsilon_{it} \quad (46)$$

$$Y_{it}(Z_{it}, M_{it}(Z_{it})) = \delta_{2t} + \gamma_t Z_{it} + \beta_t M_{it}(Z_{it}) + \eta_{it}. \quad (47)$$

for $t \in [0, 1]$. In this formulation, the total effect of Z on Y can be expressed as

$$\tau_t = \gamma_t + \alpha_t \beta_t \quad (48)$$

for all $t \in [0, 1]$. Note that here the total effect is allowed to vary across the entire range of t.

Finally, an interesting aspect of our approach which we have not previously discussed is its ability to handle missing, or non-uniformly sampled, data in the mediator variable. One of the main benefits of functional data analysis is its ability to handle such data. Though our example did not contain such data, it would be an interesting application for future work, as fMRI scans are sometimes dropped due to artifacts in the data collection at certain time points.

Acknowledgement

The author would like to thank Michael Sobel for all his help with the preparation of this manuscript, Niall Bolger for helpful comments, and Tor Wager for supplying the data. This work was patially funded by NIDA 1RC1DA028608.

References

[1] Albert, J.M. (2008). Mediation analysis via potential outcomes models. *Statistics in medicine*, **27**, 1282–1304.

[2] Amodio, D.M. and C.D. Frith (2006). Meeting of minds: the medial frontal cortex and social cognition. *Nature Reviews Neuroscience*, **7**, 268–277.

[3] Angrist, J.D. and Imbens, G.W. (1995). Two-Stage Least Squares Estimation of Average Causal Effects in Models with Variable Treatment Intensity *Journal of the American Statistical Association*, **90**, 431–442.

[4] Angrist, J.D., Imbens, G.W. and Rubin, D. B. (1996). Identification of causal effects using instrumental variables. *Journal of the American Statistical Association*, **91**, 444–455.

[5] Baron, R. M. and Kenny, D. A. (1986). The moderator-mediator variable distinction in social psychological research: Conceptual, strategic and statistical considerations. *Journal of Personality and Social Psychology*, **51**, 1173–1182.

[6] Benjamini, Y. and Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society, Series B (Methodological)*, **57**, 289–300.

[7] Bohning, D. E., Pecheny, A. P., Epstein, C. M., Speer, A. M., Vincent, D. J., Dannels, W. and George, M. (1997). Mapping transcranial magnetic stimulation (tms) fields in vivo with mri. *Neuroreport*, **8**, 2535–2538.

[8] Cardot, H., Ferraty, F. and Sarda, P. (1999). Functional linear model. *Statistics & Probability Letters*, **45**, 11–22.

[9] Davidson, R. and MacKinnon, J.G. (2010). Wild Bootstrap Tests for IV Regression. *Journal of Business & Economic Statistics*, **28**, 128–144.

[10] Efron, B. and Tibshirani, R.J. (1998). *An Introduction to the Bootstrap*. Chapman & Hall/CRC.

[11] Fan, J. and Zhang, J-T. (1999). Two-Step Estimation of Functional Linear Models with Applications to Longitudinal Data. *Journal of the Royal Statistical Society, Series B*, **62**, 303–322.

[12] Flachaire, E. (2005). Bootstrapping heteroskedastic regression models: wild bootstrap vs. pairs bootstrap. *Computational Statistics & Data Analysis*, **49**, 361–376.

[13] Grinband, J., Hirsh ,J. and Ferrera, V.P. (2006). A neural representation of categorixation uncertainty in the human brain. *Neuron*, **49**, 757–763.

[14] Hastie, T. and Tibshirani, R.J. (1993). Varying-coefficient models. *Journal of the Royal Statistical Society, Series B*, **55**, 757–796.

[15] Holland, P.W. (1988). Causal Inference, Path Analysis, and Recursive Structural Equations Models. *Sociological Methodology*, **18**, 449–484.

[16] Imai, K., Keele, L. and Yamamoto, T. (2010) Identification, inference, and sensitivity analysis for causal mediation effects. *Statistical Science*, Forthcoming.

[17] James, G. M. (2002). Generalized linear models with functional predictors. *Journal of the Royal Statistical Society, Series B*, **64**, 411–432.

[18] Jo, B. (2008). Causal inference in randomized experiments with mediational processes. *Psychological Methods*, **13**, 314–336.

[19] Lindquist, M. A. (2008). The Statistical Analysis of fMRI Data. *Statistical Science*, **23**, 439–464.

[20] 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.

[21] MacKinnon, D. P. (2008). *Introduction to statistical mediation analysis*. Mahwah, NJ: Erlbaum.

[22] Moulton, E.A., Keaser, M.L., Gullapalli, R.P. and Greenspan, J.D. (2005). Regional Intensive and Temporal Patterns of Functional MRI Activation Distinguishing Noxious and Innocuous Contact Heat. *J Neurophysiol,* **93**, 2183–2193.

[23] Müller, H.G. and StadtMüller, U. (2005). Generalized functional linear models. *Annals of Statistics*, **33**, 774–805.

[24] Neyman, J. (1923). On the Application of Probability Theory to Agricultural Experiments. Essay on Principles. Section 9, translated in *Statistical Science*, (with discussion), **5**, 465–480 1990.

[25] Pearl, J. (2004). Direct and indirect effects. In Proc. 17^{th} Conf. Uncertainty in Artificial Intelligence, San Fransisco.

[26] Petersen, M.L., Sinisi, S.E. and van der Laan, M.J. (2006). Estimation of direct causal effects. *Epidemiology*, **17**, 276–284.

[27] Ramsay, J.O., Hooker, G. and Graves, S. (2009) *Functional Data Analysis with R and MATLAB*. Springer. 2009, XII, 202 p., Softcover

[28] Ramsay, J. O. and Silverman, B. W. (2006). *Functional Data Analysis*. Second Edition. Springer, New York.

[29] Rosenbaum, P.R. and Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika*, **70**, 41–55.

[30] Robins, J.M. and Greenland S. (1992). Identifiability and exchangeability for direct and indirect effects. *Epidemiology*, **3**, 143–155.

[31] Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology*, **66**, 688–701.

[32] Rubin, D. B. (1980). Comment on "Randomization analysis of experimental data: The Fisher randomization test", by D. Basu. *Journal of the American Statistical Association*, **75**, 591–593.

[33] Rubin, D. B. (2004). Direct and indirect causal effects via potential outcomes. *Scandanavian Journal of Statistics*, **31**, 161–170.

[34] Shrout, P. E., and Bolger, N. (2002). Mediation in experimental and nonexperimental studies: New procedures and recommendations. *Psychological Methods*, **7**, 422–445.

[35] Sobel, M. E. (1982). Asymptotic confidence intervals for indirect effects in structural equation models. In S. Leinhardt (Ed.), *Sociological Methodology* (pp. 290-312). Washington DC: American Sociological Association.

[36] Sobel, M.E. (2008). Identification of Causal Parameters in Randomized Studies with Mediating Variables. *Journal of Educational and Behavioral Statistics*, **33**, 230–251.

[37] Ten Have, T.R., Joffe, M.M., Lynch, K.G., Brown, G.K., Maisto, S.A. and Beck, A.T (2007). Causal mediation analyses with rank perserving models. *Biometrics*, **63**, 926-934.

[38] VanderWeele, T.J. (2009). Marginal structural models for the estimation of direct and indirect effects. *Epidemiology*, **20**, 18–26.

[39] Wager, T.D., Davidson, M., Hughes, B., Lindquist, M. A. and Ochsner, K. (2008). Prefrontal-subcortical pathways mediating successful emotion regulation. *Neuron*, **59**, 1037–1050.

[40] Wager, T.D., Feldman Barrett, L., Bliss-Moreau, E., Lindquist, K., Duncan, S., Kober, H., Joseph, J. Davidson, M. and Mize, J. (2008). The Neuroimaging of Emotion. In Handbook of Emotions, M. Lewis, J.M. Haviland-Jones, and L.F. Barrett, Editors (pp. 249–271). Guilford Press: New York.

[41] Zhu, H., Ibrahim, J.G., Tang, N., Rowe, D.B., Hao, X., Bansal, R. and Peterson, B.S. (2007). A Statistical Analysis of Brain Morphology Using Wild Bootstrapping. *IEEE Trans. Med. Imaging*, **26**, 954–966.