14

# Call for Pragmatic Computational Psychiatry

## Integrating Computational Approaches and Risk-Prediction Models and Disposing of Causality

Martin P. Paulus, Crane Huang, and Katia M. Harlé

**Abstract**

Biological psychiatry is at an impasse. Despite several decades of intense research, few if any, biological parameters have contributed to a significant improvement in the life of a psychiatric patient. It is argued that this impasse may be a consequence of an obsessive focus on mechanisms. Alternatively, a risk-prediction framework provides a more pragmatic approach, because it aims to develop tests and measures which generate clinically useful information. Computational approaches may have an important role to play here. This chapter presents an example of a risk-prediction framework, which shows that computational approaches provide a significant predictive advantage. Future directions and challenges are highlighted.

### Biological Psychiatry: What Have You Done for Us Lately?

Biological psychiatry is in a crisis (Insel and Cuthbert 2015) for a number of different reasons. First, despite profound advances from molecular to systems neuroscience, these insights have had relatively little influence on practical psychiatry. Second, the development of new therapeutics, based on neuroscience approaches to understand the pathophysiology of these illnesses, has stalled (Insel 2012). Third, in the development of a new diagnostic classification for mental disorders (APA 2013), neuroscience had virtually no impact on contributing to the delineation and definition of the disorder categories. Fourth, there are no clinical tools for prognosis, diagnosis, and treatment

From "Computational Psychiatry: New Perspectives on Mental Illness,"
A. David Redish and Joshua A. Gordon, eds. 2016. Strüngmann Forum Reports, vol. 20,
series ed. J. Lupp. Cambridge, MA: MIT Press. ISBN 978-0-262-03542-2.

monitoring that are based on neuroscience approaches (Prata et al. 2014). Taken together, the fundamental insights into basic neuroscience have not translated into practical and clinical tools or treatment in psychiatry. Here we argue that computational approaches may play an important role in linking behavior (including emotion and cognitive processing) to neural implementations of these processes in the brain.

The lack of impact that neuroscience has had on practical psychiatry may be due to several reasons. First, one might postulate that mental health conditions, which are complex constellations of symptoms and social conditions, are fundamentally not reducible to simple biological processes. This topic deserves a thoughtful discussion, which might focus on the level of reductionism possible when observing complex clinical phenomena. Such discussion, however, is beyond the scope of this chapter.

Second, we may not have sufficiently developed technologies and approaches to map psychiatric diseases onto biological processes. This perspective is useful in generating incentives to develop new techniques in the future to advance biologically based research in psychiatry. However, the lack of progress, despite decades of increasingly sophisticated technologies, might cast doubt over the argument that it is simply a "technology problem."

Third, making biology useful for clinical psychiatry is an extremely difficult problem to solve. Given the complexity of the human brain—in terms of its amazing array of topographically organized units, which are highly interconnected, the complex orchestration of molecular events that accompany even "simple" psychological processes, and the multilevel organization that occurs from a molecular to a circuit level—this argument is hard to dispute. One would expect, however, that predictable relationships would have emerged by now between different levels of brain functioning and clinical problems.

Fourth, operational, institutional, and procedural aspects of biological research in psychiatry have not provided the appropriate environment and incentives within which biological approaches could be developed to solve clinical problems. This argument focuses on the "doing of biological psychiatry research" and might need to be addressed by leaders of funding agencies, interest groups, and research organizations.

Lastly, by focusing the search on "mechanisms" that underlie psychiatric illnesses, progress has been directed toward understanding dysfunctional processes and symptoms, rather than on clinical course and risk/protective factors. Implicit in this approach, however, is the assumption that mechanistic understanding will provide better diagnosis or treatment. We argue here that the current mechanistic viewpoint may be insufficient at this stage, and that a predictive framework may be equally fruitful to bring neuroscience to contribute to clinical psychiatry. In this context, a computational approach can provide an important framework to link behavior to neural systems processes.

From "Computational Psychiatry: New Perspectives on Mental Illness,"
A. David Redish and Joshua A. Gordon, eds. 2016. Strüngmann Forum Reports, vol. 20,
series ed. J. Lupp. Cambridge, MA: MIT Press. ISBN 978-0-262-03542-2.

## Mechanisms

The notion of a mechanism is tightly linked to causation, which can be defined as an antecedent event, condition, or characteristic that was necessary for the occurrence of the disease at the moment it occurred, given that other conditions are fixed (Rothman and Greenland 2005). Alternatively, a mechanism is, roughly speaking, a set of entities and activities that are spatially, temporally, and causally organized in such a way that they exhibit the phenomenon to be explained (Menzies 2012). Moreover, it has been highlighted that causal analyses aim to extract beliefs or probabilities that underlie observed data in both static and dynamic environments (Pearl 2010). However, causal relationships in complex systems are difficult to establish. In the context of disease and environment, Hill (1965) suggested a number of criteria in an attempt to distinguish causal from noncausal associations: strength, consistency, specificity, temporality, biological gradient (i.e., a dose-response curve), plausibility, coherence (i.e., consistency with the natural history and biology of the disease), experimental evidence, and analogy (i.e., similarities across diseases). A closer examination of examples of these criteria clearly shows that none of them are both necessary or sufficient to establish a clear causal relationship (Rothman and Greenland 2005). Moreover, there is clear evidence from carefully conducted clinical studies that causal relationships in psychiatry are difficult, if not impossible, to establish, even if many factors are considered. For example, Kendler used a propensity analysis approach to delineate covariates from causal risk factors for depression (Kendler and Gardner 2010). He concluded that dependent stressful life events, which were found to be most strongly associated with depression onset, had only a weak, if any, causal effect on the emergence of a depressive episode in the subsequent year. Further, a comprehensive analysis of the factors that influence the onset of a depressive episode shows that these factors cut across many different levels (genetic, psychological, social, economic), are highly interconnected, and differ between males and females (Kendler and Gardner 2014). These, and other results, led Kendler (2012:385) to conclude that "to develop an etiologically based nosology for psychiatric disorders is deeply problematic." Finally, in the development of increasingly sophisticated molecular approaches to understand psychiatric disorders, a silent assumption has been that one needs a more refined scale to clearly differentiate the pathophysiological processes that underlie these disorders. However, in a recent theoretical analysis of causal relationships between variables, Hoel et al. (2013) emphasized that the continued search for the "molecular cause" of a psychiatric illness may be fundamentally flawed. Specifically, they showed that one can construct interacting systems such that causal relationships emerge on a macro level but may not hold on a micro level and vice versa. Taken together, these findings suggest that at this stage a mechanistic emphasis to understanding mental disorders may delay neuroscience from making an impact for psychiatry.

From "Computational Psychiatry: New Perspectives on Mental Illness,"
A. David Redish and Joshua A. Gordon, eds. 2016. Strüngmann Forum Reports, vol. 20,
series ed. J. Lupp. Cambridge, MA: MIT Press. ISBN 978-0-262-03542-2.

We do not propose to do away completely with causal analyses, which are at the basis of mechanistic understanding of a process. Our standard statistical approaches are insufficient to clearly differentiate causal from noncausal associations. Recent attempts have been made to generate a more reliable mechanistically based quantitative theoretical framework (Pearl 2009b). Specifically, Pearl (2009a) contrasts standard statistical analyses (which aim to infer associations among variables and estimate beliefs, or probabilities of past and future events) and updates those probabilities in light of new evidence or new measurements with causal analysis. Causal analysis aims to infer not only beliefs or probabilities under static conditions, but also the dynamics of beliefs under changing conditions (e.g., induced by treatments or external interventions). Critical for this distinction, however, is to differentiate associational concepts; that is, any relationship that can be defined in terms of a joint distribution of observed variables against a causal concept, which is any relationship that cannot be defined from the distribution alone (randomization, influence, effect, confounding, "holding constant," disturbance, spurious correlation, faithfulness/stability, instrumental variables, intervention, explanation, attribution). It is important to emphasize that in psychiatry, it is very difficult to isolate causal relationships. Nevertheless, by implementing these advanced mathematical tools, we may be better able to delineate causation and, as a consequence, mechanistic frameworks for psychiatry. In the interim, however, a complementary framework may yield productive results.

## Risk-Prediction Framework

One complementary approach to the sometimes elusive search for mechanisms is to embed a program of research into a risk-prediction framework. Risk-prediction models use predictors (covariates) to estimate the absolute probability or risk that a certain outcome is present (diagnostic prediction model) or will occur within a specific time period (prognostic prediction model) in an individual with a particular predictor profile (Moons et al. 2012b). The components of risk prediction (Gerds et al. 2008) consist of (a) a sample of *n subjects*, (b) a set of *k markers* obtained for each subject, (c) an individual subject *status* at some later time *t*, which can be a scalar or vector variable, and (d) a *model* which takes the sample and markers and assigns a probability *p* of the status at time *t* for each individual.

To be useful, a prediction model must provide validated and accurate estimates of the risks; the uptake of those estimates should improve subject (self-) management and therapeutic decision making, and consequently, (relevant) individuals' outcomes and cost-effectiveness of care (Moons et al. 2012a). Risk-prediction models can be derived with many different statistical approaches. To compare them, measures of predictive performance are derived from receiver operating characteristic (ROC) methodology and probability

From "Computational Psychiatry: New Perspectives on Mental Illness,"
A. David Redish and Joshua A. Gordon, eds. 2016. Strüngmann Forum Reports, vol. 20,
series ed. J. Lupp. Cambridge, MA: MIT Press. ISBN 978-0-262-03542-2.

forecasting theory. These tools can be applied to assess single markers, multivariable regression models, and complex model selection algorithms (Gerds et al. 2008). The outcome probabilities or level of risk and other characteristics of prognostic groups are the most salient statistics for review and perhaps meta-analysis. Reclassification tables can help determine how a prognostic test affects the classification of patients into different prognostic groups, hence their treatment (Rector et al. 2012).

According to Cook (2007), one can compare the global model fit using a measure such as the Bayes's information criterion, in which lower values indicate better fit and a penalty is paid if the number of variables is increased. Moreover, one can compare general indices of calibration (e.g., the Hosmer–Lemeshow statistic, which compares the observed and predicted risk within categories) and discrimination (e.g., the c-statistic). In addition, if the overall fit for one model is better than another, but general calibration and discrimination are similar, one can assess whether the fit would be better among individuals of special interest. This would help to determine how many individuals would be reclassified in clinical risk categories and whether the new risk category is more accurate for those reclassified. Finally, one can assess utility of the risk-prediction model if it is based on an invasive or expensive biomarker, by determining whether a higher or lower estimated risk would change treatment decisions for the individual subject.

This general approach is similar to one proposed by Pencina and D'Agostino (2012), who argued that the incremental predictive value of a new marker should be based on its potential in reclassification and discrimination. In that sense, new potentially predictive (bio)markers should be assessed on their added value to existing prediction models or predictors, rather than simply being tested on their predictive ability alone (Moons et al. 2012b). The ultimate test of the effectiveness of a risk-prediction tool, like any other intervention, is a randomized clinical trial in which groups of doctors are randomized to use the tool in addition to usual care versus usual care alone (Scott and Greenberg 2010). In summary, the risk-prediction model framework has a number of advantages over a mechanistic framework: (a) a clear utilitarian approach, (b) sound statistical background, (c) a framework of iterative improvement, and (d) the ability ultimately to connect to and coexist with a mechanistic understanding of psychiatric disease. In the context of computational approaches, these aspects of the risk-prediction model framework provide clear guidance for the modeling approach: Can the underlying computational approach contribute substantially to the predictive value of the model?

Machine learning (Hastie et al. 2001) consists of a set of tools (e.g., support vector machines, random forest, recommender systems) that uses large data to understand the underlying structure (James et al. 2013). One can differentiate machine-learning tools into those that are supervised (i.e., models derived from inputs and outputs that are built for prediction or estimation) and unsupervised (i.e., models used to extract relationships and structure from

From "Computational Psychiatry: New Perspectives on Mental Illness,"
A. David Redish and Joshua A. Gordon, eds. 2016. Strüngmann Forum Reports, vol. 20,
series ed. J. Lupp. Cambridge, MA: MIT Press. ISBN 978-0-262-03542-2.

multidimensional data). Machine-learning tools have found their way into the medical field for a large number of different applications: from the prediction of healthcare services (Padman et al. 2007) to clinical predictions of the progression of Alzheimer disease (Kohannim et al. 2010; Maroco et al. 2011). Random forest is one machine-learning tool that uses predictor variables to classify members of a sample into categories (e.g., relapse or abstinent). The forest is constructed from a multitude of decision trees (Breiman 2001). Whereas a single decision tree is susceptible to noise, the average of many trees, obtained by a forest, is not, so long as the trees are uncorrelated. A random forest performs as well or better than alternative classification techniques in terms of accuracy and robustness; even in the presence of noise, the model does not overfit to a given sample (Breiman 2001). One potential downside of random forest modeling is the black box nature of the model (Strobl et al. 2009). In fact, as Breiman (2001:23) states: "a forest of trees is impenetrable as far as simple interpretations of its mechanism go." Thus, similar to other machine-learning approaches, predictive utility may, in some circumstances, come at a cost of simple mechanistic interpretations.

## Computational Approaches

While the above examples of predictive methods can be applied to all types of predictors (e.g., self-report and behavioral measures, including indicators of clinical severity and symptom types), we propose that they may be most powerful when used in combination with sophisticated inference models of beliefs and behavior. For instance, such generative models may be used at a first stage analysis to infer latent mental processes and states (associated with individual-level parameters). Such inferred states may include individuals' beliefs (e.g., about hidden reward rates of various choice options in the environment) or their decision policies (i.e., functions describing how they translate their expectations/beliefs about hidden variables into action). Recent contributions from machine learning and neuroeconomic research have highlighted several different computational approaches that can be used to infer underlying processing states from observed behavior. In our group, we have focused on two techniques, described below: optimal control and Bayesian ideal observer models.

### Optimal Control

Inverse optimal control is a computational approach used to infer an individual's reward function of a goal-directed motor task, given observed behavior. Optimal control theory has been show to be an effective computational framework to explain human movements in continuous time (Todorov and Jordan 2002). This framework is particularly valuable when examining motor

From "Computational Psychiatry: New Perspectives on Mental Illness,"
A. David Redish and Joshua A. Gordon, eds. 2016. Strüngmann Forum Reports, vol. 20,
series ed. J. Lupp. Cambridge, MA: MIT Press. ISBN 978-0-262-03542-2.

behavior; however, it can be extended readily to help understand the reward functions that drive motor behavior as part of a cognitive or affective paradigm. In this context, motor control in a goal-directed task is a dynamic process of sensorimotor integration, in which the brain takes sensory information, which includes paradigmatically specific instructions, and uses it to make continuous motor actions. Optimal control theory frames this dynamic process in a feedback control loop: the optimal controller estimates the current state at time *t*, produces a motor command based on the goal and keeps an efference copy (i.e., the expected outcome of the motor command) as the state estimator, then sends the motor command to muscles to generate the movement. The agent is thought to select actions which optimize performance on a task—the critical component of optimal control theory. In particular, the performance criterion is defined as a reward function that includes task-related performance measure and action cost. For example, in a task that instructs subjects to drive to a location A as quickly as possible, the performance measure can be the stopping distance to A; the action cost can be the accumulated effort of accelerating and decelerating controls. Individual differences are thought to emerge because subjects may have different target stopping distances and different weights to assess the ratio of the closeness to the target location over the action cost. The latter ratio defines the amount of effort one is willing to spend to achieve the intended stopping distance. Taken together, there are three components in the optimal control framework: (a) a *dynamic system* that describes how the states of the system evolve based on the action input, (b) an *action policy* that determines which action to take given the current state, and (c) a *reward function* that specifies the goal of this task (balance between goal state and action cost). A forward optimal control model generates a sequence of (optimal) actions, which maximizes the reward in the task. The goal of inverse optimal control model is to uncover the reward function assuming the optimality in observed action sequences. Proposed by Kalman (1964), inverse optimal control has been applied to study apprenticeship learning (Abbeel and Ng 2004), drivers' intention in simulated highway driving, and parking lot navigation Abbeel et al. (2008).

Using an optimal control framework (Figure 14.1) to study the behavioral processes and their dysfunction in subjects with psychiatric disorders has three main advantages. First, individuals with psychiatric disorders have been shown to have sensorimotor deficits, such as psychomotor disturbance in depressed individuals (Sobin and Sackeim 1997), which may have significant effects on the performance of effortful cognitive or affective tasks. In optimal control theory, sensory delay (i.e., the latency of an individual to respond to a visual or auditory stimulus) and motor delay (i.e., the speed with which an action plan can be carried out once the individual has selected a motor plan) can be incorporated in the dynamic system.

Second, individuals with certain psychiatric disorders may have altered reward processing; for example, individuals with depression are more sensitive

From "Computational Psychiatry: New Perspectives on Mental Illness,"
A. David Redish and Joshua A. Gordon, eds. 2016. Strüngmann Forum Reports, vol. 20,
series ed. J. Lupp. Cambridge, MA: MIT Press. ISBN 978-0-262-03542-2.

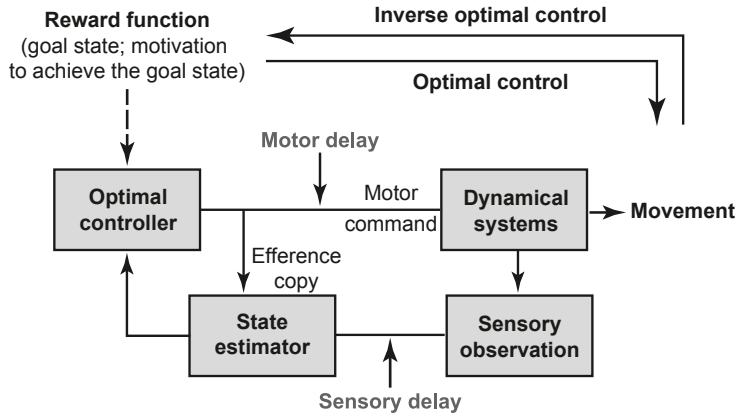


**Figure 14.1** This schema shows the basic components of the inverse optimal control framework. Each component can be estimated from a subject's motor data.

to punishment than to reward (Must et al. 2006; Eshel and Roiser 2010). The imbalance of reward and punishment sensitivity in these subjects may affect the goal state, which may differ substantially from the experimenter-instructed target state. In optimal control theory, the goal state is a parameter in the reward function that corresponds to the individual's intended state of the object in control (e.g., the position and velocity of a car in a driving task). The closer the current state is to the goal state, the higher the reward.

Third, individuals with psychiatric disorders may lack the motivation (i.e., the amount of effort to spend to achieve the subjective goal state) to perform the task (Treadway and Zald 2011; Der-Avakian and Markou 2012). In optimal control theory, motivation is also a parameter in the reward function, which measures the ratio of the weights between the accuracy to achieve goal state and the action cost in the task. The higher the motivation, the more effort one is willing to spend to achieve high accuracy toward the intended goal state. Taken together, with the appropriate experimental manipulation, we can investigate if individuals can learn and adapt their action policies to different environments by changing the dynamical system, and if their reward function will change and thus improve their performance by changing the feedback provided (e.g., reward vs. punishment).

## Bayesian Ideal Observer Models

A second computational framework which may help to extract predictive and potentially causative relationships in patients with psychiatric disorders is the

From "Computational Psychiatry: New Perspectives on Mental Illness,"
A. David Redish and Joshua A. Gordon, eds. 2016. Strüngmann Forum Reports, vol. 20,
series ed. J. Lupp. Cambridge, MA: MIT Press. ISBN 978-0-262-03542-2.

Bayesian ideal observer model or dynamic Bayesian model (DBM). DBM provides a computational framework which enables one to generate fine-grained quantification of emotion and cognitive processing as well as their interactions. The approach is to divide the observed behavior into several subprocesses, which can be submitted to test subtle hypotheses about changes in optimizing behavior. Similar to the inverse optimal control theory, DBM shares the basic assumption that changes in behavior observed in individuals with psychiatric disorders are a consequence of an altered optimization of available actions within the constraints of specific affective and cognitive processing dysfunctions. In particular, DBM is based on the notion that an individual has underlying beliefs and expectations about the situation at hand. This approach aims to quantify an individual's belief and expectation about their environment as a function of behavioral context and experienced choices and outcomes. DBM models provide a quantitative and explicit way to delineate how the brain processes complex environments, and how the breakdown of this process can contribute to the development of psychiatric disorders. Using this approach, we *infer* otherwise unknown beliefs in individuals regarding upcoming events and how such beliefs are updated based on past events experienced by the observer. This may be particularly important in a context of a risk-prediction model, when target populations exhibit very subtle or nondetectable behavioral differences on standard behavioral paradigms.

One example that shows how a simple experimental paradigm can be used to extract subtle but important cognitive control differences in healthy individuals and psychiatric subjects is the application of DBM to inhibitory control using the stop signal task. Specifically, Yu and colleagues used DBM to capture behavioral adjustments on a trial-by-trial basis of stopping behavior (Shenoy et al. 2010; Shenoy and Yu 2011; Ide et al. 2013). This model is based on the assumption that an individual updates the prior probability of encountering Stop trials, $P$(stop), on a trial-by-trial basis, based on trial history; it adjusts decision policy as a function of $P$(stop), with systematic consequences for Go response times and Stop accuracy in the upcoming trial. An optimal response when an individual assumes that it is more likely to encounter a Stop trial—that has a higher $P$(stop)—is to slow one's response latency (i.e., exhibit a slower Go response time), which would result in a higher likelihood of correctly stopping on a Stop trial. This adjustment has been shown in two different experiments in healthy subjects (Ide et al. 2013; Harlé et al. 2014). To model the trial-by-trial adjustment of prior expectations, we used a Bayesian hidden Markov model adapted from the DBM (Yu and Cohen 2009; Ide et al. 2013). The model makes the following assumptions about subjects' internal beliefs regarding task structure: on each trial $k$, there is a hidden probability $r_k$ of observing a Stop signal ($s_k = 1$ for Stop trial) and a probability $1 - r_k$ of observing a Go trial ($s_k = 0$); $r_k$ is the same as $r_{k-1}$ with probability $\alpha$ and is resampled from a prior beta distribution $p_0(r)$ with probability $1 - \alpha$. The predictive probability of trial

From "Computational Psychiatry: New Perspectives on Mental Illness,"
A. David Redish and Joshua A. Gordon, eds. 2016. Strüngmann Forum Reports, vol. 20,
series ed. J. Lupp. Cambridge, MA: MIT Press. ISBN 978-0-262-03542-2.

$k$ being a Stop trial, $P_k(\text{stop}) = P(s_k=1 \mid \mathbf{s}_{k-1})$, where $\mathbf{s}_k = (s_1, ..., s_k)$ is a vector of all past trial outcomes, 1 for Stop trials and 0 Go trials, can be computed as:

$$\begin{aligned} P\left(s_k = 1 | s_{k-1}\right) &= \int P\left(s_k = 1 | r_k\right) p\left(r_k | s_{k-1}\right) dr_k \\ &= \int r_k \, p\left(r_k | s_{k-1}\right) dr_k = \left\langle r_k | s_{k-1} \right\rangle. \end{aligned} \tag{14.1}$$

Predictive probability of seeing a Stop trial, $P_k(\text{stop})$, is the mean of the predictive distribution $p(r_k|\mathbf{s}_{k-1})$, which, by marginalizing over the uncertainty of whether $r_k$ has changed from the last trial, becomes a mixture of the previous posterior distribution and a fixed prior distribution, with $\alpha$ and $1 - \alpha$ acting as the mixing coefficients, respectively:

$$p\left(r_k | s_{k-1}\right) = \alpha \, p\left(r_{k-1} | s_{k-1}\right) + \left(1-\alpha\right) p_0\left(r_k\right). \tag{14.2}$$

Posterior distribution over Stop trial frequency is updated according to Bayes's rule:

$$p\left(r_k | s_k\right) \alpha P\left(s_k | r_k\right) p\left(r_k | s_{k-1}\right). \tag{14.3}$$

The DBM model further assumes a positive linear relationship between trial-wise $P$(stop) and reaction times at the individual level. That is, on a given trial, the higher the expected likelihood of encountering a Stop signal, the more a person should slow down to avoid making a Go error. For each parameter setting (i.e., each pair of alpha and the prior distribution mean), the corresponding $P$(stop) sequence can be inferred, and linear regression can be used to determine the optimum parameter values providing the strongest correlation coefficient or R square coefficient between $P$(stop) and reaction times. In our previous work, we have found that subjects' prior resampling rates to be best captured with alpha values between .6 and .8 (Shenoy and Yu 2011; Ide et al. 2013; Harlé et al. 2014).

## An Example Study: Predicting the Emergence of Problem Stimulant Use

While significant executive deficits have been demonstrated in chronic stimulant dependence (Salo et al. 2002; Monterosso et al. 2005; Hester et al. 2007; Tabibnia et al. 2011), only subtle behavioral impairments in error monitoring and inhibitory control have been observed in individuals at risk for stimulant dependence (Colzato et al. 2007; Reske et al. 2011). Thus, in the following example, we applied DBM to the analysis of event-related functional magnetic resonance imaging (fMRI) data associated with baseline inhibitory function during a stop signal task to predict clinical status three years later. Previously, healthy volunteers (Ide et al. 2013) and individuals at risk for stimulant use disorder (Harlé et al. 2014) were shown to adapt their response strategy in

From "Computational Psychiatry: New Perspectives on Mental Illness,"
A. David Redish and Joshua A. Gordon, eds. 2016. Strüngmann Forum Reports, vol. 20,
series ed. J. Lupp. Cambridge, MA: MIT Press. ISBN 978-0-262-03542-2.

this inhibitory control paradigm. Thus, we hypothesized that computational models might help identify neural substrates that contribute to subtle inhibitory deficits. Such computational neural variables were hypothesized to perform significantly better than other variables, such as noncomputational task-based brain activity and clinical measures (e.g., cumulative drug use), in predicting long-term clinical status (Harlé et al. 2014).

We recruited occasional stimulant users (OSUs) from the student population of different local universities. OSUs were defined primarily as having (a) at least two off-prescription uses of cocaine or prescription stimulants (amphetamines and/or methylphenidate) over the past six months and (b) no evidence of lifetime stimulant dependence. Participants completed a baseline interview session to evaluate clinical diagnoses and determine current patterns of drug use as well as a neuroimaging session that examined brain and behavior responses during decision making; they completed a stop signal task while being scanned. We were able to follow these OSUs for three years, after which they completed another standardized interview (phone or in-person) which examined the extent of drug use over the three-year interim period (using the SSAGA II). Two groups of interest were identified: problem stimulant users (PSUs) and desisted stimulant users (DSUs). PSUs were *a priori* defined by (a) continued stimulant use since baseline interview and (b) endorsement of 2+ symptoms of DSM-IV amphetamine and/or cocaine abuse and/or dependence criteria occurring together 6+ contiguous months since the initial visit. DSUs endorsed (a) no 6-month periods with 1+ stimulant uses and (b) no symptoms of interim stimulant abuse or dependence.

Using a split-sample approach, we first identified potential predictive neural regions with voxel-wise robust logistic regressions to predict three-year follow-up status (coded 1 = PSU vs. 0 = DSU) in a randomly selected "training" subset of our sample. The remaining "test" subset was used to assess the relative predictive power of the activation clusters identified with the training sample by using random forest analysis (Breiman 2001). In this study, we ran three random forest analyses, each with a distinct set of baseline variables to compare the overall performance of (a) drug-use measures (total uses of stimulants, cocaine, and marijuana, based on self-report), (b) categorical fMRI regressors (task-based contrasts such as Stop vs. Go, Stop Success vs. Stop Error), and (b) Bayesian/computational fMRI regressors, respectively. To construct those regressors, we first convolved three types of trials (Go, Stop Success/SS, and Stop Error/SE) with a canonical hemodynamic response function in a general linear model (GLM). Each of these predictors were entered both as linear regressors and parametrically modulated by trial-level $P$(stop) estimates. This model allowed us to isolate neural activations associated with both trial type alone (i.e., categorical regressor) and $P$(stop). Thus, after deconvolution, this model included six task regressors. Three were categorical: Go, SS, SE. Three were model-based parametric: Go × $P_k$(stop), SS × $P_k$(stop), SE × $P_k$(stop). A second GLM was created with trial-wise Bayesian signed prediction error

From "Computational Psychiatry: New Perspectives on Mental Illness,"
A. David Redish and Joshua A. Gordon, eds. 2016. Strüngmann Forum Reports, vol. 20,
series ed. J. Lupp. Cambridge, MA: MIT Press. ISBN 978-0-262-03542-2.

(SPE), defined as Outcome – *P*(stop), and unsigned prediction error (UPE), defined as |Outcome – *P*(stop)|, included as parametric regressors of interest. Individual subjects' percent signal change (%SC) scaled beta weight values for five regressors; contrasts of interest from these two GLM models were extracted and used as independent variables in the prediction analyses. The categorical regressors included two contrasts: (a) (Stop – Go), that is, (SS + SE)/2 – Go and (b) (SE – SS). The Bayesian regressors included three computational predictors: (a) *P*(stop), that is, ½*Go × $P_k$(stop) + ¼ *SS × $P_k$(stop), + ¼*SE × $P_k$(stop); (b) UPE; and (c) SPE. For full description of these first-level fMRI analyses, see Harlé et al. (2014).

Based on logistic regressions in the training sample, predictors in the full model included activations extracted from 21 ROIs identified with robust logistic regressions, including three ROIs for trial type-independent *P*(stop) activation, six ROIs associated with Bayesian UPE activation—UPE: Outcome – *P*(stop)—and twelve ROIs associated with SPE activation—SPE: |Outcome – *P*(stop)|. Based on random forest analyses in the test sample, we found that:

1. no variable met criteria for inclusion in the drug-use model, which had an overall accuracy of 52%;
2. only one variable met criteria for inclusion in the fMRI categorical predictor model (i.e., SE – SS contrast activation in the rostral anterior cingulate cortex), but with an overall accuracy of 64%, which was not significantly different statistically from the no-predictor model based on response rate alone; and
3. four variables met criteria for inclusion in the fMRI computational predictor model, including UPE activation in right thalamus as well as SPE activation in right anterior insula/inferior frontal gyrus, in the right superior medial prefrontal cortex/dorsal anterior cingulate cortex (BA32), and in right caudate (BA25).

Notably, this final model yielded an overall accuracy of 74%, which represents a statistically significant improvement in accuracy from the model based on response rate alone.

The utility of the computational approach can be most easily displayed using a Bayesian nomogram (Figure 14.2). The vertical axis of the left-hand side of the nomogram shows the prior probability of developing problem use, or the proportion of the total sample that showed problem use. The vertical axis of right-hand side shows the posterior probability of problem use given a positive or a negative test result, respectively. The vertical axis in the center displays the positive or negative likelihood ratio, which is the most important characteristic of a test in terms of linking the knowledge before applying the test to the knowledge once the test has been conducted and found to be either positive or negative. In this instance, we used the random forest model as the basis for the testing procedure. The upper and lower brackets around the central estimate represent the 95% confidence interval of the post-test probability, providing

From "Computational Psychiatry: New Perspectives on Mental Illness,"
A. David Redish and Joshua A. Gordon, eds. 2016. Strüngmann Forum Reports, vol. 20,
series ed. J. Lupp. Cambridge, MA: MIT Press. ISBN 978-0-262-03542-2.

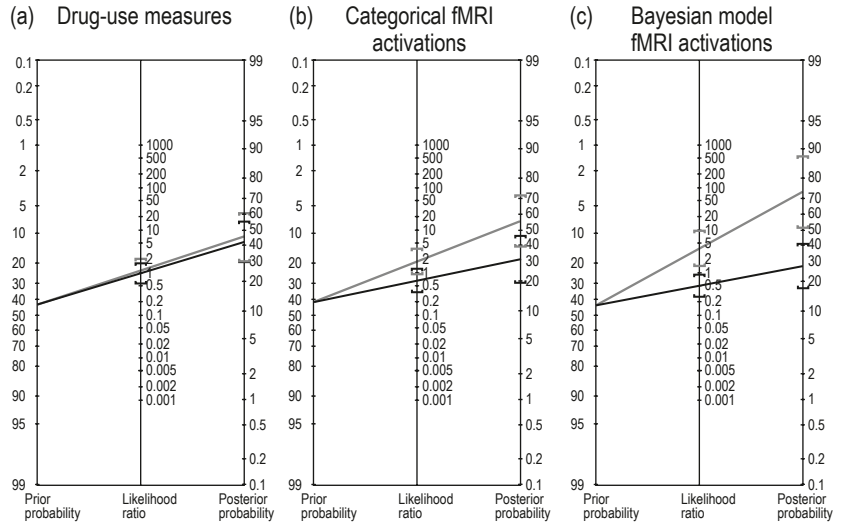


**Figure 14.2** Bayesian nomogram for (a) drug-use measures, (b) categorical fMRI activation measures, and (c) Bayesian model-based activation measures. The positive likelihood ratio (gray) and negative likelihood ratio (black) show that the test based on the Bayesian ideal observer model clearly provide the best separation. For example, based on a base rate of approximately 57%, a positive test indicated a 74% chance of becoming a problem user, whereas a negative test reduces the chance to approximately 28%.

a graphical means of indicating whether the test measurably improved our knowledge; that is, yielded a higher or lower post-test probability without the confidence interval including the pretest probability. Thus, when the 95% confidence intervals do not intersect, positive and negative tests are statistically significantly different.

For all four computational predictors, larger neural responses negatively correlated with Bayesian prediction errors were associated with a higher likelihood to be categorized in the PSU group three years later. Specifically, for every standardized unit increase in UPE deactivation in right thalamus, one was about three times as likely to develop a future stimulant-use disorder (odds ratio = 3.45, $p < .05$). In addition, an individual was two to three times as likely to be categorized in the PSU group for every standardized unit increase in SPE deactivation in medial prefrontal cortex/anterior cingulate cortex (odds ratio = 2.44, $p < .05$), anterior insula/ inferior frontal gyrus (odds ratio = 3.19, $p < .05$), and caudate (odds ratio = 3.02, $p < .05$). To summarize visually the relative predictive power of the three predictive models considered, we used bootstrapped robust logistic regressions to produce cumulative ROC curves associated with each added layer of predictor types. As seen in Figure 14.3, the computational predictors added in the last layer (black line) significantly increased accuracy, including both sensitivity and specificity.

From "Computational Psychiatry: New Perspectives on Mental Illness,"
A. David Redish and Joshua A. Gordon, eds. 2016. Strüngmann Forum Reports, vol. 20,
series ed. J. Lupp. Cambridge, MA: MIT Press. ISBN 978-0-262-03542-2.

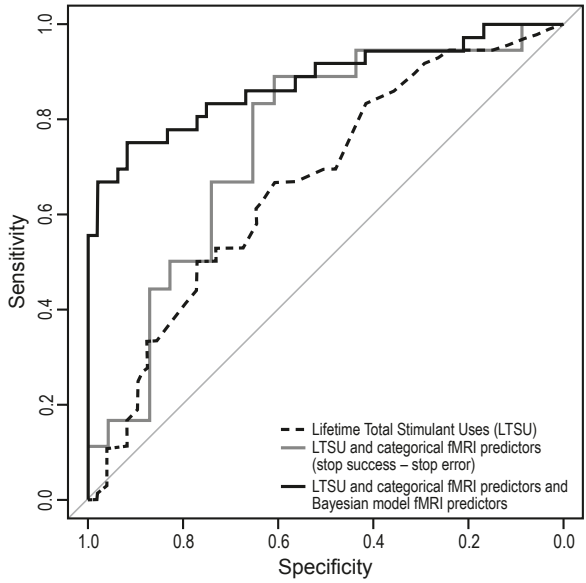


**Figure 14.3** Receiver operator curve for three different predictor models. Lifetime stimulant use is indicated by the dotted black line; lifetime stimulant use plus fMRI and behavioral measures in gray; and lifetime stimulant use, fMRI, and Bayesian model parameters are shown in black. The Bayesian ideal observer model parameters add significantly to increased sensitivity and specificity of the model.

In this study we sought to determine whether the combination of functional neuroimaging and computational approaches to behavior are able to generate predictions that can help to determine whether an individual will progress to problem use. Using the combination of Bayesian ideal observer model, stop signal task as a measure of inhibitory control, and fMRI imaging, we show that those individuals who demonstrated greater neural processing, related to a Bayesian prediction error, are more likely to develop a future stimulant-use disorder over the subsequent three years. For this, we used a Bayesian ideal observer model to quantify an individual's belief about the likelihood of an upcoming Stop trial (i.e., the person's probabilistic expectation of having to mount an inhibitory response during a stop signal task). Importantly, these data were collected in individuals at risk for stimulant-use disorder three years prior to the assessment of the outcome (i.e., whether the subject would progress to problem use or desist using). Cross-validated robust regression and random forest analyses showed that neural responses associated with Bayesian model-inferred prediction errors (representing the trial-wise discrepancy between expectation of a Stop trial and actual trial outcome) in right anterior cingulate cortex, anterior insula, caudate, and thalamus most robustly predicted three-year clinical status (i.e., meeting criteria for stimulant-use disorder vs. desisted-use status). These computational neural variables significantly contributed

From "Computational Psychiatry: New Perspectives on Mental Illness,"
A. David Redish and Joshua A. Gordon, eds. 2016. Strüngmann Forum Reports, vol. 20,
series ed. J. Lupp. Cambridge, MA: MIT Press. ISBN 978-0-262-03542-2.

predictive validity above the base rate, which was not the case for other baseline predictors *a priori* thought to be promising, such as reported lifetime drug use or non-model-based neural predictors. Therefore, this study shows that, in principle, a Bayesian cognitive model applied to an event-related neural activity can be used to predict long-term clinical outcome. Taken together, these results are consistent with the notion that the combination of functional neuroimaging and computational modeling can provide better predictions of future clinical states.

## Future Directions

This example is one among a series of emerging studies in neurology and psychiatry that use machine-learning approaches in the context of risk-prediction models to generate individual-level predictions (Perlis 2013; Moradi et al. 2015). The emphasis of computational approach, up to now, has been on improving a mechanistic understanding of behavior in complex situations. For example, temporal difference models (Schultz et al. 1997) provide a clear and convincing framework for the acquisition of reward (Daw and Touretzky 2002) and aversive learning (Iordanova 2009) in both animals (Schultz and Dickinson 2000) and humans (Klein-Flugge et al. 2011). The extension to Bayesian models has been based on the notion that humans utilize not only information about experienced averages but also about the underlying distribution; that is, the degree of uncertainty associated with the experience (Behrens et al. 2007). The use of computational models provides a powerful technique to disambiguate processes that result in an observable behavior and can thus be used to make inferences about processes which constrain behavior in a way that is observed in psychiatric populations (Huys et al. 2011).

There are, however, several caveats that one must keep in mind when applying these models in psychiatry. First, our diagnostic descriptions of patients are at best initial phenomenological approximations of heterogeneous subgroups of individuals (Insel and Cuthbert 2015). Consequently, behavioral dysfunctions are likely to result from different underlying pathologies associated with different computational processes. In other words, it is unlikely that individuals with anxiety or depression will show a uniform computational dysfunction.

Second, psychiatric illnesses are a mixture of long-range dysfunctions, best captured by trait variables and momentary dysregulation, which are assessed using state measures. Moreover, there may be different stages of psychiatric illnesses based on the recovery from the illness process itself. For example, individuals who show substance-use disorder might undergo prolonged recovery of function, which may result in changes of the computational process that guides the decision making. Thus, it will be important to examine psychiatric populations at different stages of illness to better understand the dynamics of the underlying process dysfunction.

From "Computational Psychiatry: New Perspectives on Mental Illness,"
A. David Redish and Joshua A. Gordon, eds. 2016. Strüngmann Forum Reports, vol. 20,
series ed. J. Lupp. Cambridge, MA: MIT Press. ISBN 978-0-262-03542-2.

Third, Paulus (2007) has previously proposed that decision making is a homeostatic process closely related to the physiological state of the body. As a consequence, computational processes that underlie the selection of an option might be sensitively affected by the individual's body state. For example, decision making focused on the selection of food items are highly dependent on the satiety state of the individual (Haase et al. 2009). Therefore, it should be clear that computational processes may need to be examined in the context of different motivational states. Taken together, the use of computational models in psychiatry to explain the underlying mechanism of behavior is promising, but at an early stage and will require many future studies to delineate some of the issues raised above.

Alternatively, computational approaches may have a more immediate impact on psychiatry by providing better prediction models. In this context, the value of the computational approach is to modify the receiver operator statistic that determines where to set a cut off for a positive or negative test, or to improve the likelihood ratio of the test being developed. The goal here would be to use underlying belief updating models to improve the prediction of future behavior or future clinical outcomes. This approach does not rely on a particular disease category (i.e., whether an individual has major depressive disorder, dysthymia, or bipolar depression). Instead, the risk-prediction model framework aims to exploit individual differences to make better predictions. However, several issues need to be taken into account:

Clinically relevant and robust predictions require large data sets. Currently, most studies, with a few exceptions (Whelan et al. 2014), are based on relatively small samples. Thus, it will be important to collect data sets that are based on "real" patient populations of sufficient size to be able to make robust predictions.

In addition, it is unclear at which level one is best able to delineate a causal pathway to the pathology in psychiatric illnesses. The Research Domain Criteria approach (Insel et al. 2010) relies on the assumption that more basic molecular levels will eventually result in a better causal prediction of the emergence and maintenance of psychiatric illnesses. This assumption may, however, be deeply flawed, as greater clarity at the molecular level might not necessarily yield stronger causal relationships (Hoel et al. 2013). In fact, the presence of a romantic relationship in an adolescent's life resulted in the single strongest predictor of the emergence of binge drinking (Whelan et al. 2014), whereas genetic and neuroimaging markers were only weakly predictive. The use of a risk-prediction framework will act as an arbiter of what is the best information to be clinically useful. Moreover, it may also act as a pointer to show us where to carve nature at its joints. Computational approaches might well play an important role here, but we are still early in this endeavor.

From "Computational Psychiatry: New Perspectives on Mental Illness,"
A. David Redish and Joshua A. Gordon, eds. 2016. Strüngmann Forum Reports, vol. 20,
series ed. J. Lupp. Cambridge, MA: MIT Press. ISBN 978-0-262-03542-2.