10

A Novel Framework for Improving Psychiatric Diagnostic Nosology

Shelly B. Flagel, Daniel S. Pine, Susanne E. Ahmari,
Michael B. First, Karl J. Friston,
Christoph Mathys, A. David Redish,
Katharina Schmack, Jordan W. Smoller,
and Anita Thapar

Abstract

This chapter proposes a new framework for diagnostic nosology based on Bayesian principles. This novel integrative framework builds upon and improves the current diagnostic system in psychiatry. Instead of starting from the assumption that a diagnosis describes a specific unitary dysfunction that causes a set of symptoms, it is assumed that the underlying disease causes the clinician to make a diagnosis. Thus, unlike the current diagnostic system, this framework treats both symptoms and diagnostic classification as consequences of the underlying pathophysiology. Comorbidities are therefore easily incorporated into the framework and inform, rather than hinder, the diagnostic process. Further, the proposed framework provides a bridge—which did not previously exist—that links putative constructs related to pathophysiology (e.g., RDoC domains) and clinical diagnoses (e.g., DSM categories) related to signs and symptoms. The model is flexible; it is expandable and collapsible, and can integrate a diverse array of data at multiple levels. Crucially, this novel framework explicitly provides an iterative approach, updating and selecting the best model, based on the highest-quality available evidence at any point. In fact, the scheme can, in principle, automatically ignore data that is not relevant or informative to the diagnostic trajectory. Finally, the proposed framework can account for and incorporate the longitudinal course of an illness. This

Group photos (top left to bottom right) Shelly Flagel, Danny Pine, David Redish, Jordan Smoller, Susanne Ahmari, Karl Friston, Michael First, Katharina Schmack, Danny Pine, Katharina Schmack, group discussion, Susanne Ahmari, Anita Thapar, Christoph Mathys, Michael First, David Redish, Karl Friston, Shelly Flagel, Jordan Smoller, Anita Thapar, Christoph Mathys

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.

chapter details the theoretical basis for this framework and provides clinical examples to illustrate its utility and application. Multiple iterations of this framework will be required based on available information. It is hoped that, with time, the framework will enhance our understanding of individual differences in brain function and behavior and ultimately improve treatment outcomes in psychiatry.

Introduction

Nosology is defined as the branch of medicine that addresses the classification of disease (see First, this volume). In psychiatry, such classifications capture the ways in which patients present to clinicians and are meant to assist mental health professionals in providing optimal clinical care. The current fundamental approach to nosology in psychiatry is to classify disorders as categorical syndromic clusters of signs, symptoms, and potentially laboratory findings, as outlined in the Diagnostic and Statistical Manual of Mental Disorders (DSM). Clearly, this approach has had a major positive impact on clinical practice and has led to substantial improvements in research and improved understanding of some neural mechanisms underlying dysfunction. Nevertheless, the clinician's reliance on the DSM categorical descriptive approach to diagnosis has had significant limitations in terms of assisting clinicians in treatment selection and prediction of prognosis. Although it was clearly hoped that the DSM system would reflect the etiology or the pathophysiological processes that underlie the disorders, it has become equally clear that this is not the case (Hyman 2010). Although relatively little was known in this regard when DSM-III emerged more than thirty years ago (APA 1980), there is little to no mention of neurobiological processes, even in the more recent version, DSM-5 (APA 2013), despite great advances in psychiatric neuroscience. As a result, consensus has emerged regarding the need to go beyond the categorical descriptive approach of DSM, with the hope of improving outcome prediction and treatment response for individual patients.

The Research Domain Criteria (RDoC), sponsored by the U.S. National Institute of Mental Health (NIMH), were established to provide "a framework for creating research classifications that reflect functional dimensions stemming from translational research on genes, circuits, and behavior" (Insel and Cuthbert 2009:989). The goal of RDoC was to shift researchers toward a focus on dysregulated neurobiological systems, as the organizing principle for delineating dysfunction. For example, constructs in RDoC, such as "cognitive control," emerge from neuroscience research that links functions in neural circuits to measures of information processing obtained in the laboratory (e.g., Morris and Cuthbert 2012). Although RDoC has not, to date, led to great improvements in diagnostic nosology, the foundation it provides for linking nosology and neuroscience is critical for advancing the field in both research and clinical domains.

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.

Clearly, both dimensional and categorical approaches such as RDoC and DSM offer complementary advantages, and both approaches to nosology are likely to be invaluable for many years to come. The problem, however, is that there is currently no link between the neurobiological systems that form the basis of the fundamental RDoC domains and the symptoms that form the basis for the syndromes that comprise the DSM. As a result, neuroscience remains far removed from clinical decision making. Although there are advantageous features of the DSM and RDoC which, if combined, could potentially improve and advance clinical treatment and research related to mental illness, there is currently no means to bridge these two systems. That is, there is no framework that will allow clinicians or researchers to link domains of psychological or neurobiological function to existing DSM diagnostic categories. Further, there is no mechanism in place to relate RDoC domains or DSM constructs to measures of clinical utility (e.g., prognosis, effective treatment, cost). Current approaches to categorization are also deficient in their ability to account for comorbid diagnoses and the longitudinal evolution or trajectory of disorders. Here, we introduce an integrative framework, based on Bayesian principles, that builds upon the current systems. It integrates the process of clinical decision making and the process by which individual differences in brain function give rise to individual differences in behavior. The ultimate goal of this framework is to improve diagnoses and treatment outcomes in psychiatry.

Utilizing an Integrative Framework to Improve the Diagnostic Process

Computational (or theoretical) neuroscience likely has much to offer in terms of developing a mechanistically informed nosological structure; however, it may also have a useful role in providing a formal probabilistic structure to the ensuing diagnostic process in and of itself. In short, our approach treats symptoms, signs, and clinical decisions about a patient as observable consequences of latent constructs (such as deficits in "cognitive control"), which emerge from hidden physiological states (see Figure 10.1). Such a framework can be used progressively to reduce uncertainty about clinical decision making by guiding the clinician to the most informative questions to ask or diagnostic tests to perform. Further, inherent in this framework is the ability to compare models of pathophysiological and psychopathological dynamics, and their causes and consequences, to quantify the relative confidence in differential diagnoses. This, in principle, could include the clinicians' more intuitive (expert) inferences, which could be recorded as diagnostic scores (see Friston, this volume), much like scores from clinical questionnaires. Prior to presenting the technical components that constitute this framework, we first introduce some key concepts and terminology.

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.

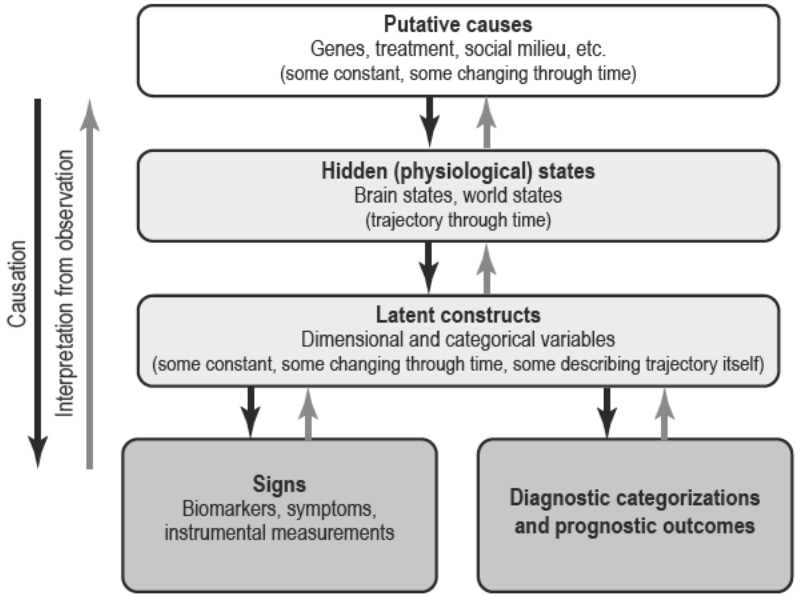

Figure 10.1 Conceptual drawing of the Bayesian Integrative Framework. In this novel integrated nosological framework, diagnoses are placed with symptoms and other observations that arise from underlying pathophysiological causes. Specific causes in the subject or the world create physical and physiological states which are understood through latent constructs that produce observations observable in the world. Thus, causation is assumed to flow from top to bottom in this figure. Causes can be constant (e.g., the genome of a subject) or changeable over time (e.g., time since an externally driven trauma). The underlying physical and physiological states are assumed to be unobservable and to be changing through time. These states are understood through dimensional and categorical latent constructs, such as cognitive control or negative affect. These latent constructs produce observable signs, such as symptoms, measures on instruments, or detectable neural signals. Diagnostic categorizations are additional observations based on clinical expertise. Similarly, prognostic outcomes are observable categorizations (e.g., recovery or relapse).

Normative versus Process Models

When thinking about how the brain works, there is an important distinction between models in computational science that is indicated by the adjectives "normative" and "process." This distinction is important both conceptually and practically. In brief, *normative models* describe how an "optimal" system would work given the goals; in other words, normative models describe "what" the brain is trying to do. In contrast, *process models* are fundamentally about the mechanisms, thus describing "how" it is done. Although these two types of models are often placed in contrast to each other, they actually represent two sides of the same coin. Briefly, every process model implies the existence of a normative model, asking what the optimal solution is given the underlying

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.

limitations of that process. Similarly, every normative model has implications for the potential processes that can produce that optimal solution. For example, many normative decision models do not take into account the time it takes to compute a predicted expected outcome, nor do they take into account the limited perceptual abilities humans have. However, once one has specified the processes for computing expected outcomes or the limitations of one's perception, one can ask about the normative model, given those limitations.

More technically, normative models rest on the assumption that the behavior at hand can be cast as an optimization process, where the states or parameters of the normative model optimize a well-defined function. In contrast, a process model specifies the algorithmic and implementational details of how the objective is attained. Process models can be formulated in terms of putative neuronal processes. Indeed, the utility of process models is that they can be used as observations or statistical models of observed responses such that, when fitted to data, their parameters associate biological processes with a functional role (Boly et al. 2011). Thus, process models have the useful property of characterizing neurophysiological responses in terms of well-defined computations.

The standard process model implicit in most nosologies assumes that a particular entity exists, as indicated by a diagnostic term, which causes symptoms. For example, "schizophrenia" is viewed as an external reality in a patient, which causes hallucinations and other manifestations of the illness. Here, we adopt a different view of diagnosis, considering it an observation based on the judgment of the clinician. From this perspective, continuously distributed latent variables cause a diagnosis. As described below, a latent variable is an unobservable state or parameter that is inferred or hypothesized based on the observable states or parameters. For example, a neuronal failure of perceptual inference causes hallucinations, delusions, and a concomitant diagnosis of schizophrenia. When a clinician arrives at a diagnosis, this represents one important piece of observational data that can be placed in the context of other observations. Thus, the diagnosis does not cause symptoms, but rather reflects the impact of multiple underlying processes. Using Bayesian inference, probabilistic statements can be made about the nature of these underlying processes. Moreover, as information accumulates, these statements can be progressively updated in a process that, over time, may improve clinical prognosis. This Bayesian process model is based on the notion that a diagnosis does not cause a disease process—the disease process causes the diagnosis—and its accompanying symptoms and signs (Figure 10.1).

Use of a Bayesian Model to Improve Diagnostic Nosology

Bayesian inference allows reasoning about uncertain quantities (i.e., latent variables) according to the rules of probability theory. In other words, Bayesian

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.

inference is the extension of deductive logic (i.e., reasoning about quantities which are certain) to inductive reasoning (reasoning about quantities which are probabilistic). One can show (Cox 1946) that any violation of the rules of probability (or, equivalently, of inductive reasoning or Bayesian inference, which are different names for the same thing) entails a violation of common sense. We can therefore use Bayesian models to relate the cause-effect inferences above.

Latent Variables/Latent Constructs

Latent or hidden variables are states or parameters that are not directly observed, but can be inferred by inverting a model of how observations depend on them. To give a simple example, consider the interpretation of a diagnostic test of HIV—a single observation in this system. We interpret the relationship between the presence of HIV infection (*A*) and the results of the test (*B*) based on Bayes's theorem:

$$p(A \mid B) = \frac{p(B \mid A) \cdot p(A)}{p(B \mid A) \cdot p(A) + p(B \mid \neg A) \cdot p(\neg A)}, \tag{10.1}$$

where *A* and *B* refer to statements that can be true or false, and ¬*A* is the negation of *A*. Applied to clinical data, *A* could be, "The patient is HIV positive," and *B* could be "The patient's HIV test is positive." $p(A)$ denotes the probability that statement *A* is true, and $p(A \mid B)$ denotes the probability that *A* is true given that *B* is true. Assuming that the infection rate for the demographic profile of the patient is 1% (i.e., $p(A) = 0.01$ and $p(\neg A) = 0.99$), that the true positive rate of the test is 95% (i.e., $p(B \mid A) = 0.95$), and that the false positive rate is 2% (i.e., $p(B \mid \neg A) = 0.02$), Bayes's theorem tells us that, given an initial positive test, the probability that the patient is HIV positive is 32%:

$$\frac{0.95 \cdot 0.01}{0.95 \cdot 0.01 + 0.02 \cdot 0.99} = 0.32. \tag{10.2}$$

The usefulness of Bayes's theorem derives from the fact that it allows us to quantify the effect on the probability that *A* is true, after establishing that *B* is true. Moreover, as observations accrue, we also are able to incorporate each emerging observation into this accrual process. Before taking any initial observation on *B* into account, based on the infection rate in the relevant population, $p(A)$ is 0.01 and $p(\neg A)$ is 0.99. This is called the *prior probability distribution*, often referred to simply as the *prior*, because it is the prior belief before taking any additional observations into account. After observing *B*, Bayes's theorem then gives us $p(A \mid B) = 0.32$ and $p(\neg A \mid B) = 0.68$, which is the *posterior probability distribution*, or simply *posterior*, because it is the (updated) distribution after taking the observation into account.

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.

The latent variable here is the presence of HIV infection in the blood of the patient (A or $\neg A$). It cannot be observed directly; rather, it has to be inferred by inverting (i.e., applying Bayes's theorem to) a probabilistic model of how HIV status leads to test outcomes. Evidence regarding the latent variable can be accumulated by repeated observation. Since a probability of 32% for positive HIV status is a poor basis for a treatment decision, we can choose to apply the test again. Assuming the test comes back positive again, we now have a probability of 96%:

$$\frac{0.95 \cdot 0.32}{0.95 \cdot 0.32 + 0.02 \cdot 0.68} = 0.96. \tag{10.3}$$

In this calculation, we simply had to replace 0.01, our original estimate based only on demographics, with 0.32, our estimate that followed from our initial observation of B. Moreover, in this calculation, we replace 0.99 with 0.68, similarly reflecting the influence of the first test. The probability of the patient being HIV positive had risen from 0.01 to 0.32, and conversely, the probability of his or her being HIV negative shrank from 0.99 to 0.68. In Bayesian terms, the *posterior* after the first observation is the new *prior* before the second observation.

Information about latent variables cannot only be accumulated by repeated observations of the same kind, but also by integrating information from different sources. If we assume that after the first test was positive, a second, perhaps more expensive, test with true positive rate 99% and false negative rate 0.5% was used and came back positive, then the probability of HIV positive status after the second test would have been 99%:

$$\frac{0.99 \cdot 0.32}{0.99 \cdot 0.32 + 0.005 \cdot 0.68} = 0.99. \tag{10.4}$$

These examples illustrate that if we have models of how latent variables lead to observations, we can use Bayesian belief updating to infer the status of latent variables and base treatment decisions and prognoses on those inferences. Indeed, in the proposed integrative framework, we are using Bayesian principles to do just this for psychiatry.

Bayesian Model Comparison

Bayesian inference can also be used to score the goodness of formal generative models. For example, one model might describe the evolution of a particular psychopathology as cyclical, leading to a prediction of periodically fluctuating clinical observations, whereas another might describe the evolution of the same psychopathology as linear, leading to a prediction of linearly progressing clinical observations. Given actual observations, the goodness of each model

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.

can be scored in terms of its expected predictive power, which has two aspects: the ability to explain existing data and the ability to generalize to new data. These two requirements are jointly quantified by one number: the model evidence. When applied to a given dataset, each competing model has a certain amount of evidence for it. Formally, the model evidence is calculated by marginalizing (i.e., taking a probability-weighted sum of) the model likelihood over all possible latent variable values:

$$\begin{aligned} &\text{Mode evidence} = p(\text{Observations} \mid \text{Model}) \\ &= \int p(\text{Observations} \mid \text{Variables})\, p(\text{Variables})\ \text{dVariables}, \end{aligned} \tag{10.5}$$

where dVariables denotes integration (i.e., summation) over the whole variable space, p(Observations | Variables) is the likelihood, and p(Variables) weights the likelihood by the prior probability of the variables taking a particular value. While the resulting number is not interpretable in isolation, the *ratio* of two model evidences is the *Bayes's factor* which indicates the *relative* quality of two models. This is because the Bayes's factor is what relates the prior odds (i.e., the ratio of the probability of one model to the probability of the other before making any observations) to the posterior odds (i.e., the same ratio after making observations):

$$\text{Posterior odds} = \text{Bayes's factor} \cdot \text{prior odds}, \tag{10.6}$$

or more formally:

$$\frac{p(\text{Model 1} \mid \text{Observations})}{p(\text{Model 2} \mid \text{Observations})} = \frac{p(\text{Observations} \mid \text{Model 1})}{p(\text{Observations} \mid \text{Model 2})} \cdot \frac{p(\text{Model 1})}{p(\text{Model 2})}. \tag{10.7}$$

A Bayes's factor greater than 1 indicates more evidence for Model 1 than for Model 2, while a Bayes's factor of less than 1 indicates more evidence for Model 2 than for Model 1.

It is important to note that Bayesian model comparison does not decide which model is correct; it quantifies the evidence supporting each of the candidate models. In doing this, it automatically accounts for both the accuracy and complexity afforded by each model. One can show (Penny et al. 2004) that:

$$\text{Model evidence} = \text{Accuracy} - \text{Complexity}, \tag{10.8}$$

where accuracy and complexity both have formal definitions. Introducing additional latent variables increases the complexity of a model, but these additions might yet improve model evidence if the complexity increase is outweighed by an increase in accuracy. In general, there will be a peak in model evidence

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.

for a certain number of latent variables, after which adding more complexity is no longer warranted.[1]

Bayesian Integrative Framework

Although we can only observe (e.g., symptoms, measurements), we assume that these observations arise from an underlying (unobservable) reality. This underlying reality includes many diverse factors (e.g., the social milieu, the brain state of the subject, epigenetics) that are too complex to measure directly. We define each of these variables to be a dimension within a space. A given subject occupies some point in this very high-dimensional space.[2] Since these variables are assumed to be unobservable, we assume that there are a set of *latent variables* or *constructs* that capture the most important aspects of this underlying reality. A given subject at a given moment in time occupies some position within this space of latent variables that is defined by dimensional constructs. Over time, the subject traces a trajectory through that multidimensional space. Because we do not actually observe the latent variables directly, we use Bayesian analysis methods (see below) to derive a probability distribution over a position at a given time, and over the trajectories a subject is taking through that space. An example of trajectories through this dimensional space is given by Friston (this volume), who provides a simulation of how such a framework could work.

These latent variables reflect dimensional constructs that arise from our understanding of neuroscience, psychology, and other sciences. For example, one might wish to quantify the attention abilities of a subject, how the subject arbitrates between deliberation and habit-based decision making, whether the subject's behavior reflects problems with emotions or impulse control, how reactive a subject's amygdala is to emotional stimuli, etc. These latent variables are, by definition, dimensional, incomplete, and mutable with new discoveries. Furthermore, in this framework, some subjects at particular locations within this state-space of latent variables manifest clinical observations that would lead a clinician to place the subjects into one or more diagnostic categories at any time. Indeed, generating a categorical diagnosis from latent variables that correspond to the dimensions allows us to model comorbidity in terms of underlying pathophysiological or psychopathological dimensions, and to assign a unique mapping to diagnostic categories. Because diagnostic categories reflect (potentially overlapping) areas of state-space, a single point in state-space may correspond to several diagnostic categories. Of course, the actual location of the subject in state-space is unknown, but a probabilistic distribution over

1 Popular model scoring measures such as the Akaike information criterion or the Bayesian information criterion are approximations to the model evidence which use the exact accuracy term but approximate the complexity term because this term is harder to compute.

2 The state-space is the set of all possible values the latent variables can take; given the large number of potential variables, the state-space is described here as "high dimensional."

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.

points in state-space can be inferred from the symptoms shown by the subject. Because the multiple diagnostic categories overlap in some, but not other, areas of state-space, comorbidity can increase the accuracy of this state-space prediction.

Of note, there had been hope that the diagnoses in the DSM and its ilk would reflect specific latent variables—that all of the subjects diagnosed with a disorder, such as obsessive-compulsive disorder (OCD), would have a similar underlying neuropsychological dysfunction. However, decades of research have established this hypothesis to be incorrect. Importantly, and perhaps unfortunately, a diagnosis made using taxonomies such as the DSM is a single measurement, while the reality is much more complex. For example, the relationship between a diagnosis and neuropsychological dysfunction shows both equifinality (a particular symptom can arise from multiple dysfunctions) and multifinality (a particular dysfunction can generate multiple symptoms; see below for further discussion). The Bayesian Integrative Framework we propose captures this complexity by separating the diagnoses from the underlying latent variables/constructs. For this reason, we treat the diagnosis as an observation reflecting underlying latent variables, with a diagnosis being a clinician's measurement of the patient within a scheme such as DSM or the International Classification of Diseases (ICD).

The levels of this framework are connected by probabilistic (Bayesian) reasoning (Figure 10.2). We can measure the probability of a symptom S_i arising from each latent variable (or combination of latent variables) LV_j as $P(S_i \mid LV_j)$ or $P(S_i \mid LV_j, LV_k, \ldots)$. Using Bayes's rule (see above and Mathys, this volume), we can invert these probabilities to infer the probability of a latent variable having a specific value, given the observations $P(LV_j \mid S_1, S_2, \ldots S_n)$. What this means is that we can use observations to predict which values we expect the latent variables to take. For example, a diagnosis of OCD may only be partially predictive of a dysfunction in the balance between goal-directed and habit-based decision-making systems (Gillan et al. 2011). Thus, if a clinician diagnoses a patient as having OCD, we can use this framework to infer the probability of having a dysfunction in the balance between goal-directed and habit-based decision-making systems.

To be thorough, we should describe the probability of each observation S as arising from a trajectory through the space of latent variables $P(S_i \mid \text{path of } LV_j, \text{path of } LV_k, \ldots)$, and again, we can invert this probabilistic model to describe the likelihood of following a path through the space of latent variables from the observations $P(\text{path of } LV_j \mid S_1, S_2, \ldots)$. Thus, for example, many patients presenting to a clinician with a diagnosis of major depressive disorder will have suffered from a series of relapsing depressed episodes. The number and spacing of these episodes may be more informative for the likelihood of treatment successes than the particular constellation of symptoms that any given patient expresses when seen by the clinician (e.g., Tundo et al. 2015). The relapsing and remitting nature of these episodes forms a trajectory through an

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.

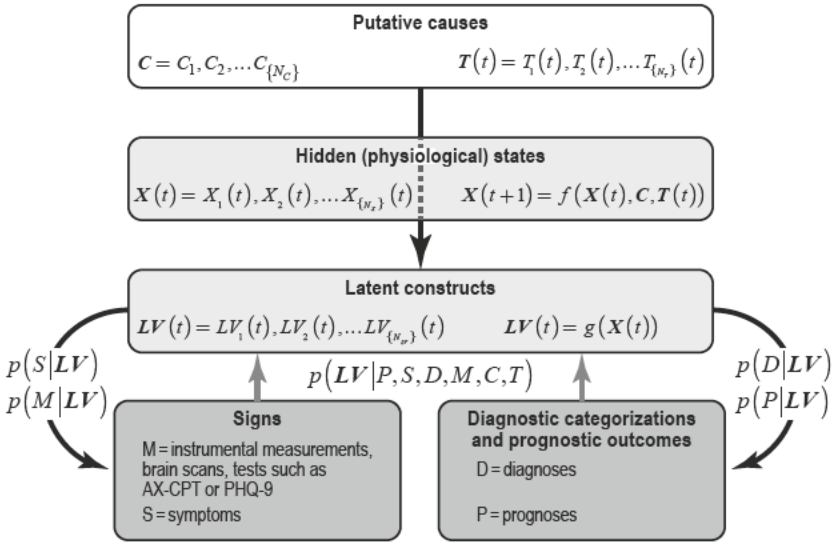

Figure 10.2 Mathematical formulation of the Bayesian Integrative Framework. As in Figure 10.1, the variables are connected through Bayesian inference, and causation is assumed to flow from top to bottom. Thus, putative causes probabilistically cause hidden physiological states, which probabilistically cause latent constructs, which probabilistically produce signs and outcomes. Because we cannot directly observe the hidden physical/physiological states, we will directly use the probability of obtaining a latent construct given the putative causes. Because we are using a Bayesian framework that can be inverted using Bayes's rule, we can calculate the probability of obtaining a latent construct given the putative causes and the observed signs and diagnoses. Similarly, we can calculate the probability of observing a given sign or diagnosis from the derived latent constructs. Thus, for example, we could measure signs (such as PHQ-9 scores or brain scans), calculate how that changes the probability distribution across latent constructs, and then calculate how that changes the probability of a given diagnosis or prognosis. AX-CPT: continuous performance task; PHQ-9: patient health questionnaire; *LV*: latent variable; *X*: hidden physiological states; *C*: putative causes; $p(\ldots)$: probability function; *t:* time point; *T*: treatment.

underlying space of both symptoms (such as reports of dysphoria) and latent variables (such as negative cognitive schema). In addition, we can incorporate the probabilities arising from static underlying causes such as genetics.

The interaction of probabilities arising from this scheme means that latent variables that are informative provide tight probabilities to observations, and latent variables that are not informative do not. Thus, we do not need to commit to specific latent variables at any particular point in time; rather we can allow the latent variables to change with basic science discoveries and the development of new conceptual frameworks in the coming years.

The following important points need to be considered in this framework:

1. The position of the patient within the latent variable space is probabilistic. That is, we do not say that the patient has followed a specific

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.

trajectory or will follow a specific trajectory through that latent variable space, but rather that there is a probability that any trajectory is the real one.
2. There are multiple levels of understanding and data included. The recognition that the diagnosis of the clinician is actually a measurement, not a latent variable in itself, is important because it allows the incorporation of information from neuroscience that derives from dimensional conceptual frameworks such as RDoC. Importantly, there is no limit to the set of conceptual frameworks that can be so incorporated.
3. This framework allows one to move both up (from observations to constructs) and down (from constructs to observations).

Because the framework is fundamentally dynamic, it includes mechanisms to incorporate new variables and to remove variables that are not informative. Thus, when developing the framework, the following points are important:

1. We do not need to rebuild this framework from scratch. Because the framework is modifiable, we can use the current DSM categories and insights available from the RDoC project (or elsewhere) as starting points and modify them as needed. In the Bayesian terminology, these are our priors.
2. This framework can incorporate new measurements, whether they are new tasks, new symptomologies, or new diagnostic schemata. It can be updated based on new evidence to enhance its explanatory value.
3. This framework allows one to use Bayesian model comparisons to make decisions about which models and latent variables to include.

In this framework, the patient is described as taking a trajectory through the multidimensional space of latent variables. What those latent variables are will change as new findings from basic science emerge. It might be useful to think of this as a blurry set of possible trajectories: some of them may be very sharp and clear, whereas others less so.

In this framework, one could take a diagnosis made by a clinician at a given time, use that to predict a probability distribution across latent constructs, and then use that probability distribution across latent constructs to predict future outcomes, including symptoms, task-related observations, and diagnoses. In general, the strength of this proposal is that it allows a set of observations (e.g., task performance, measurements) to be used to predict a distribution across constructs which can then be used to predict a future diagnosis. In this scheme, *prediction* arises from the future trajectory. The trajectory describes how subjects typically pass through the paths in the space of latent variables. This feature will allow a diagnostician to better infer future outcomes on a case by case basis. *Treatment*, in this case, is about bending the curve; it is about changing the probabilities along the future trajectories, which changes 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.

latent variables, which changes the observations. An example of how treatment can bend trajectories can be found in the simulations in Friston (this volume).

Because trajectories are very high-dimensional (i.e., they describe the time course of many latent constructs) and complex, we can also view future trajectories as *prognoses*. Prognosis is really a set of predicted observations (diagnoses or other outcomes) at some future time. Therefore, we could capture the concept of prognosis by including both present and predicted future outcomes in our analyses. Thus, it is not necessary to specify a prognosis explicitly within this framework. However, it may be useful to do so. Within this framework, a prognosis is an outcome/observation. One could, for example, create a new outcome/symptom/observation of interest (e.g., "will relapse/will not relapse"), which could then be predicted from the distribution across constructs. The decision to treat or not could then be determined by examining how the prognosis is conditional on treatment.

Phases of Application

Some of the pragmatic advantages of the Bayesian Integrative Framework can be appreciated in terms of the phases of its application:

Phase I: Construction of the Framework. This step could begin with an expert consensus that provides the clinician with the optimized model in a computerized form. As part of this phase, all of the empirical observations (e.g., signs, symptoms, treatment history and diagnoses at various time points) would be used to build a generative model that evolves over time and describes a trajectory. Of course, the process of generating such a model is quite important, and the details of this process would need to be explicated. However, the process of refining the model is more important than the process of generating the initial model. As a result, our discussion here devotes space primarily to discussing refinement of the model. Broadly conceptualized, to create this initial model, the number of latent variables, causal inputs, and associated functions that best describe outcomes in terms of symptom profiles and other measurements, such as the clinician's diagnosis, would be optimized using standard Bayesian inversion schemes and Bayesian model selection (see above). Thus, this phase uses existing data, or priors, to optimize the model per se. It should also be noted, as described below (see Phase Ia), that theories specific to a particular disorder or cluster of symptoms can be tested with this framework and used to further refine and enhance the model.

Phase II: Application. This step is the application for the clinician. After having selected the best model, the posterior relative to the parameters (i.e., weights) of that model can be used as priors to estimate the posterior constructs and hidden states that provide the best explanation for a patient's symptom profile and associated signs (e.g., neuroimaging, or clinical) and measures. Thus

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.

the clinician would use all available data to make such estimates. This might only include data from a mental-status examination and history. Alternatively, such clinical data could be accompanied by data from neuropsychological tests and brain-imaging experiments. Regardless of the content, the process of estimation relies on the same procedures; the posterior distribution of latent causes of the patient's symptoms can now be used to create a posterior predictive density over differential diagnoses (e.g., DSM) at the current time and, crucially, in the future. Furthermore, one could simulate probabilistic responses to therapeutic interventions in terms of (probabilistic) trajectories over future (diagnostic or therapeutic) outcomes. Taken together, the knowledge gathered in Phase I can be integrated into an application in Phase II for a clinician to use to get from observations (symptoms, diagnoses) to predicted trajectories of future outcomes (prognoses). The predicted trajectories would be reported as probabilities over the multiple possible future outcomes.

Phase III: Refinement. This step would again emerge from a consensus of experts. The generative model could be refined by using all the clinician's observations made during the application of the model. The accumulated data from application could be assimilated using Bayesian belief updating (as described above) to improve the parameter estimates and model selection. This recursive procedure could be iterated indefinitely, providing an increasingly efficient description of "good diagnostic practice" and therapeutic outcomes. In addition, one would have the opportunity to include (or eliminate) constructs and hidden states using Bayesian model selection. Ultimately, the trajectories of hidden states underlying disease progression (or its resolution) may acquire increasingly mechanistic details as our understanding of pathophysiology accumulates.

Examples of Application of the Bayesian Integrative Framework

To illustrate how the proposed integrative nosological framework could enable the integration of findings from psychiatric research into the clinician's diagnostic act in a well-defined and quantifiable manner, we present three specific examples. Using the phases of the application described above, we explain how complex empirical results could inform the clinician's prognostic predictions and treatment decisions for individual patients presenting with attenuated psychotic symptoms, posttraumatic stress disorder (PTSD), or OCD.

Attenuated Psychotic Symptoms

Attenuated psychotic symptoms can, in some cases, herald late onset schizophrenia. A diagnosis for "attenuated psychosis syndrome" is now included in the DSM-5 (Research Appendix) as a condition warranting further

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.

investigation. Epidemiological studies suggest, however, that less than 40% of the individuals with attenuated psychotic symptoms will develop full-blown schizophrenia within five years after diagnosis (Fusar-Poli et al. 2012a, 2013a). Despite accumulating evidence about risk and protective factors for the transition from attenuated psychotic symptoms into schizophrenia (Fusar-Poli et al. 2013b), prognostic predictions in clinical practice are still very imprecise.

Phase I: Construction of the Framework

Empirical observations from various time points are used to construct a generative model of the underlying static and dynamic unobservable causes, states, and constructs (Figure 10.3). These empirical observations can be specific *symptoms* (e.g., ideas of reference or delusions), *functional outcomes* (e.g., social or occupational functioning), or *diagnoses* (e.g., attenuated psychosis syndrome or schizophrenia) assessed at various time points by longitudinal studies. These empirical observations can further comprise any *sign* or *biomarker* that has been associated with the transition from attenuated psychotic symptoms into psychotic disorder. For instance, a positive family history for psychosis is a strong predictor for transition (Seidman et al. 2010; Thompson et al. 2011), indicating that genetic risk might be a putative cause of the development of schizophrenia.

Moreover, the trajectory from attenuated psychotic symptoms into psychotic disorder has been associated with reduced cortical volume in prefrontal, cingulate, and insular regions (Smieskova et al. 2010), as well as with a range of neurocognitive deficits and impaired social perception (Fusar-Poli et al. 2012b). This points to a role for constructs, such as "cognitive control" or "perception and understanding of others," in the development of schizophrenia. The generative model can also accommodate observed influences of *treatment* on the trajectory toward schizophrenia, such as the reduction of transition rates into schizophrenia by psychosocial interventions (Preti and Cella 2010; van der Gaag et al. 2013).

Phase Ia: Refinement and Testing of the Hypothesis

After comparing models with different observed and unobservable variables, the best generative model is then selected. To illustrate the potential power of model selection, let us assume that the selected generative model for the trajectory of attenuated psychotic symptoms into schizophrenia includes both the constructs "perception and understanding of others" and "cognitive control." From a pragmatic perspective, the use of this refined model would yield more precise predictions regarding the trajectory of an individual patient (see Phase II: Application). Most importantly, however, the models derived in the integrative framework are not agnostic with respect to the mechanisms underlying

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.

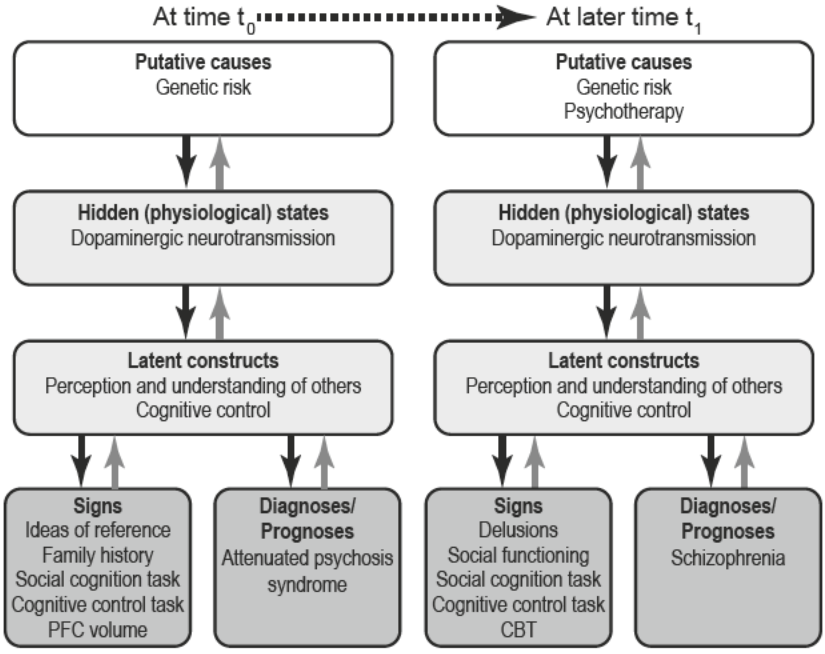

Figure 10.3 Generative model describing the trajectory from attenuated psychotic symptoms to schizophrenia. Putative causes can be either constant (genetic risk) or time-varying (psychotherapeutic interventions) and are assumed to generate neurophysiological states (altered dopaminergic neurotransmission). These neurophysiological states are linked to constructs of mental function (perception and understanding of others and cognitive control). The unobservable causes, neurophysiological states, and latent constructs ($LV(t)$) generate time-varying observations in the form of symptoms (e.g., ideas of reference), signs (e.g., reported family history of psychosis), and biomarkers (e.g., performance in a social cognition task). By model inversion, the observations can be used to infer the unobservable causes, states, and constructs to enable clinical predictions: Will this patient develop schizophrenia? How should we treat this patient? (See also Figure 10.4.) They can also provide mechanistic insight into the pathophysiology of psychotic symptoms: What is psychosis? PFC: prefrontal cortex; CBT: cognitive behavioral therapy.

the trajectory to schizophrenia. For instance, selecting between models which include either "perception and understanding of others" or "cognitive control," or both, would provide formal tests of hypotheses regarding the neuropsychology of psychotic symptoms.

Phase II: Application

Let us imagine that a 20-year-old patient presents with intermittent ideas of reference with mostly preserved reality testing. The clinician can now take the patient's history and use the reported symptoms to make a probabilistic

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.

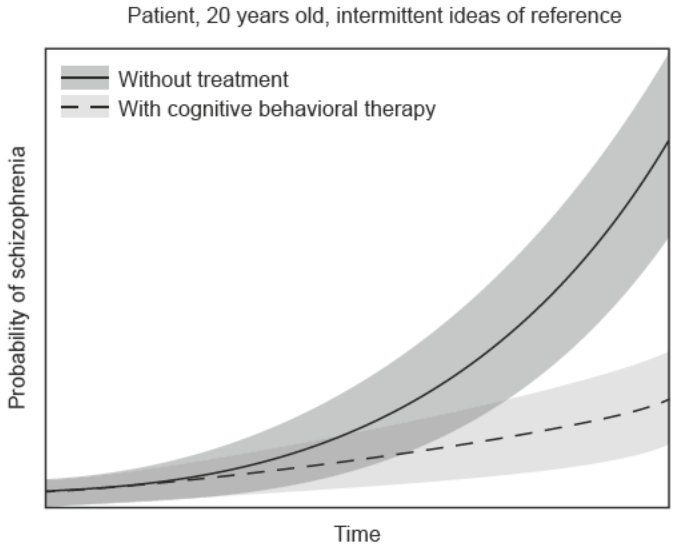

Figure 10.4 Example of a probabilistic trajectory for an individual patient as predicted by the generative model. In the application phase, the clinician can enter clinical observations (e.g., severity of ideas of reference, family history of psychosis) into the refined and estimated generative model, and get time courses of posterior distributions over diagnoses (e.g., schizophrenia) or symptoms (e.g., social functioning). The lines represent the mean of these posterior distributions; the shaded areas depict confidence intervals. Entering additional clinical observations (e.g., task scores in a social cognition task or prefrontal cortical volumes measured by an MRI scan) would result in a narrowing of the confidence intervals equivalent to an increased precision of the predicted probabilistic time courses. Further, the model could simulate probabilistic time courses in response to a therapeutic intervention (e.g., cognitive behavioral therapy).

prediction about the patients' trajectory into full-blown schizophrenia. Alternatively, it would also be possible to make a prediction about the trajectory of any sign, symptom, or functional outcome (e.g., social functioning). Moreover, a probabilistic prediction for a trajectory without treatment can be compared to the probabilistic prediction for a trajectory with treatment (Figure 10.4). The clinician can also decide to obtain further information for the patient (e.g., a structural MRI scan to measure prefrontal cortical volume or a neurocognitive test to quantify cognitive control). This information can then be used along with the reported symptoms to increase the precision of the probabilistic predictions about the trajectory.

Phase III: Refinement

Given that the proposed integrative framework can incorporate different types and amounts of data, the generative model can be continuously refined using

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.

new empirical evidence. For instance, a recent large-scale genome-wide association study (GWAS) established an association between schizophrenia and several polymorphisms in genes related to immune function (Schizophrenia Working Group of the Psychiatric Genomics Consortium 2014). To accommodate this finding, the generative model could be extended by the observed genotypes and the neurophysiological state "immune function." The refined model would enable more precise predictions about the trajectory of attenuated psychotic symptoms in individual patients that have been genotyped for the respective genetic variants. Most notably, however, by concurrently considering the new neurophysiological state ("immune function") together with other neurophysiological states ("dopaminergic neurotransmission"), novel mechanistic hypothesis regarding the etiology of psychotic symptoms could be generated and tested.

Posttraumatic Stress Disorder

PTSD represents another especially complex disorder (Shay 1994; Kessler et al. 1995; Cantor 2005). It has interesting temporal components that are more easily accommodated within this framework than in standard practice.

Phase I: Construction of the Framework

First, we need to take what is known about PTSD and build a framework based on these known interactions. Within our integrative framework, we could include measurements of symptoms or signs (e.g., number of recalls, emotionality of recalls, where the recalls occurred, lack of sleep) as well as the DSM categorizations as observations. We would also want to include the time since the occurrence of the traumatic event as an observation. Importantly, this integrative framework is able to add in other factors, which may or may not be related to one another. For example, we could factor in a preexisting cause based on hippocampal size. Data from twin studies show that soldiers with PTSD as well as their non-trauma-exposed twin (who does not have PTSD) have smaller hippocampi than soldiers without PTSD and their twins (Gilbertson et al. 2002). This can be factored in by including an additional constant, Hipp, reflecting hippocampal size (Figure 10.5), allowing us to ask whether the addition of this constant (hippocampal size) changes the predicted probabilities of different latent variables.

Phase Ia: Refinement and Testing Theories

There are three classes of theories regarding dysfunction in PTSD: it entails (a) an *encoding* error (Brown and Kulik 1977), (b) a *decoding* error (Nadel and Jacobs 1996), and (c) a *recovery* error (Redish 2013). The latter can be separated into a lack of normal posttraumatic recovery or a worsening of symptoms.

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.

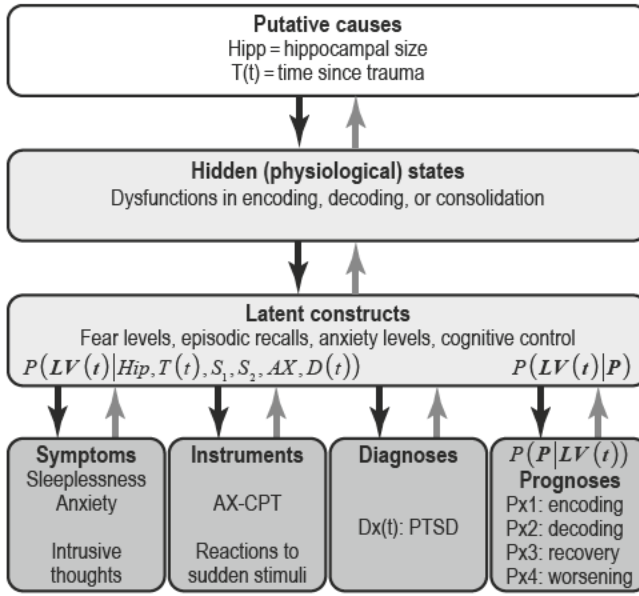

Figure 10.5 Generative model describing a patient presenting with PTSD symptoms. A set of hypothesized and explanatory latent constructs ($LV(t)$) are assumed to vary over time and can be predicted from putative causes, including constant causes, such as hippocampal size (Hipp), and temporally changing causes, such as time since trauma ($T(t)$) as well as from observations such as symptoms (S), measures on instruments (AX-CPT: continuous performance task), and clinical diagnoses (D). The latent constructs will show a progression through time, likely following one of the four hypothesized theories: an encoding deficit, a decoding deficit, a lack of recovery deficit, or a worsening (anti-recovery) trajectory. Prognoses are effectively a categorization of these trajectories. The integrative framework permits theories to be tested and prognoses to be predicted, allowing it to be used in fundamental science (e.g., causes of PTSD) as well as clinical science (e.g., PTSD treatments).

All of these hypotheses suggest differences across the patient's trajectory after the trauma. The *encoding* hypothesis suggests that the trauma is encoded differently in the patients who develop PTSD compared to those who do not. For example, a patient who develops PTSD after a bus explosion may have encoded this event with more emotional valence than one who does not develop PTSD, and this is perhaps related to more stress-related cortisol in their brain at the time of the trauma (LeDoux 1996; Jacobs and Nadel 1998; Shors 2004). The *decoding* hypothesis suggests differences in how the patient recalls the traumatic event later. That is, the patient may have a generalization deficit, in which they do not successfully identify the circumstances that indicate danger (Nadel and Jacobs 1996; Jovanovic and Ressler 2010). For example, a soldier may overreact to a surprising touch on the shoulder or may be unable to sit in a crowded restaurant after returning home from a combat zone because they incorrectly retrieve danger signals (Shay 1994). The *recovery* hypothesis

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.

suggests that trauma is encoded similarly between people who develop PTSD and those who do not; however, people who do not develop PTSD recover from their trauma differently than people who do (with PTSD patients possibly getting worse over time). For example, in "normal" trajectories of memory function, memories are initially encoded episodically with a strong "you-are-there" component, which creates a mental time travel component during recall. Subsequently, with time, storytelling, and sleep, those memories become semantic narratives and are decoupled from the episodic mental time travel (Nadel and Moscovitch 1997; Squire 2004; Redish 2013). For PTSD patients, however, the memory continues to be stored episodically.

These three theories can be differentiated through prospective research that charts the trajectory of the PTSD symptoms and their associated effect on functioning over time. Imagine comparing changes in symptoms over time after trauma between two groups of individuals, one that eventually develops PTSD and another that does not. The encoding hypothesis suggests that there would be large differences in reactions to the trauma and associated symptoms virtually immediately after the trauma. In contrast, the decoding hypothesis suggests that the differences in symptoms would not manifest immediately after the trauma, but would emerge soon thereafter and continue over time, but with little change. Finally, the recovery hypothesis suggests that non-PTSD subjects would change more over time than PTSD subjects. We could test all three of these hypotheses using the proposed integrative framework.

Phase II: Application

The advantage of this framework is that it does not assume that PTSD is a single phenomenon represented by one of those three theories—it is possible that any given patient may have an encoding error, a decoding error, or a recovery error. The Bayesian Integrative Framework provides probabilities of each of these underlying dysfunctions from the set of observed symptoms.

Presumably, each of these dysfunctions will require different treatments. As the scientific community is refining and testing the theories (Phase Ia), it will also be necessary to determine how future outcomes (symptoms, diagnoses, prognoses) are affected by different treatments. From the probabilities of each of these trajectories, it should be possible to identify which treatments would be best suited to which patient on an individualized basis.

Phase III: Refinement

As stated above, one of the major advantages of this integrated framework is its inherent flexibility. As new neurophysiological and neuropsychological measures become available, it is easy to incorporate those new observations into the Bayesian equations. For example, we could add another neurobiological variable of functional connectivity. Georgopoulos et al. (2010) and his team

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.

have reported connectivity differences as detected with magnetoencephalography (MEG) in patients categorized with PTSD relative to those who do not have the disorder. These measurements could be integrated as new instruments in Figure 10.5. With our integrative framework, we could ask whether or not repeated measures of these connectivity differences allow clearer identification of the latent variables or of the patient's trajectory through that space of latent variables. They have, for example, also found that these measurements change over time, with more changes in controls than in veterans with PTSD (Anders et al. 2015). This example is illustrative of how advances in psychiatric neuroscience research will continuously inform and improve the results of this model in years to come.

Obsessive-Compulsive Disorder

For several reasons, OCD provides another useful test case for our proposed integrative framework. First, compared to other neuropsychiatric disorders, OCD has excellent convergence in results from neuroimaging studies, demonstrating consistency in identified brain regions with abnormal structure and function (orbitofrontal cortex, anterior cingulate cortex, striatum, and anterior thalamus) across imaging modalities (e.g., structural MRI, fMRI, PET, DTI) and research sites (Baxter et al. 1988; Swedo et al. 1989; Rauch et al. 1994; Alptekin et al. 2001; Menzies et al. 2008). This contributes to multiple types of reliable observations that can be plugged into the model, including neuroimaging findings as well as symptoms and DSM diagnoses. This, in turn, may lead to greater power due to the generation of smaller confidence intervals. In addition, with OCD we have a clear theoretical hypothesis regarding a potential set of latent variables that may have importance in driving OCD symptoms; this will be described further below in Phase Ia.

Phase I: Construction of the Framework

In Phase I, as described in the previous two examples, the proposed Bayesian Integrative Framework can be used to assemble empirical observations about OCD from both clinical experience (e.g., DSM-based diagnoses) and the clinical literature (e.g., genetic risk factors, performance on neurocognitive tasks, treatment outcomes) to build a generative model. As delineated in Figure 10.6, these observations could include a broad array of data such as:

- clinical measurement of symptom presence (obsessions, compulsions, anxiety levels, tics),
- symptom types (e.g., contamination obsessions/compulsions, doubt obsessions/checking compulsions),
- symptom levels (as measured by reliable instruments such as YBOCS, HAM-A, HAM-D; see Figure 10.6),

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.

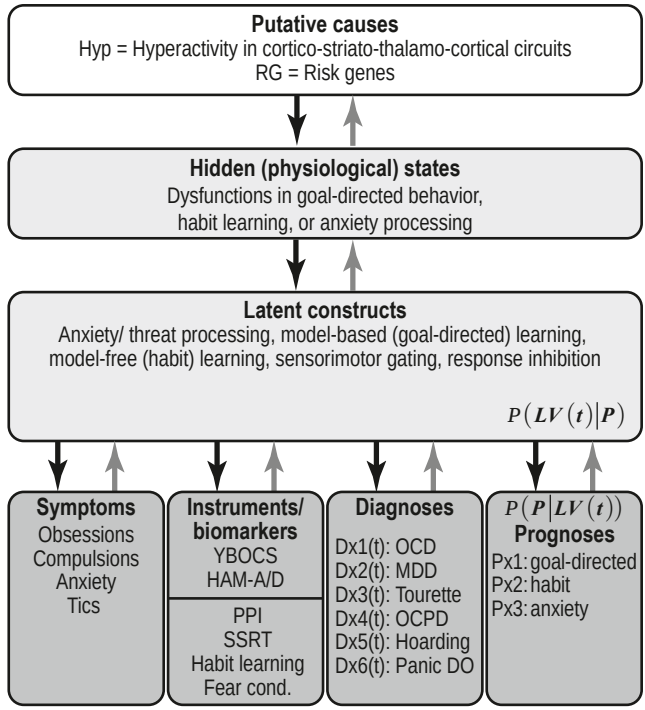

Figure 10.6 Generative model describing a patient presenting with OCD symptoms. As for PTSD, a set of hypothesized and explanatory latent constructs ($LV(t)$) are assumed to vary over time and can be predicted from putative causes—including constant causes, such as risk genes (RG), and temporally changing causes, such as hyperactivity in cortico-striato-thalamo-cortical circuits (Hyp)—and from observations (e.g., symptoms, measures on instruments, biomarkers, and clinical diagnoses). Latent constructs will show a progression through time, likely following one of the three hypothesized theories: deficits in goal-directed behavior, habit learning, or anxiety processing/expression. Prognoses are effectively a categorization of these trajectories. The integrative framework allows theories to be tested and prognoses to be predicted, making it useful to address basic issues in fundamental science (e.g., causes of OCD) as well as clinical science (e.g., effective treatments for OCD). YBOCS: Yale-Brown obsessive-compulsive scale; HAM-A: Hamilton anxiety rating scale; HAM-D: Hamilton depression scale; PPI: prepulse inhibition; SSRT: stop signal reaction time; Fear cond.: fear conditioning; MDD: major depressive disorder; OCPD: obsessive-compulsive personality disorder; Panic DO: panic disorder.

- level of insight,
- presence of comorbidities (e.g., major depressive disorder, tic disorder/Tourette syndrome, obsessive-compulsive personality disorder, hoarding disorder),
- response to both pharmacotherapy and psychotherapy treatments,
- performance on measures of neurocognitive functions implicated in OCD (e.g., habit-learning tasks, prepulse inhibition),

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.

- fear conditioning,
- stop signal reaction time tasks, and
- putative causes (e.g., hyperactivity in cortico-striato-thalamo-cortical circuits that can be observed with both PET and fMRI; risk genes).

Importantly, these variables can be weighted according to the strength of the available evidence so that consistently replicated findings would make a greater contribution to the model. For example, despite the fact that no GWAS study to date has identified strong genome-wide candidates for OCD risk genes (Stewart et al. 2013; Mattheisen et al. 2015), genetic association studies have implicated an association between the glutamate transporter, SLC1A1, and OCD (Bloch and Pittenger 2010; Wu et al. 2012). Having an *SLC1A1* risk allele would therefore impact the model. In addition, as previously discussed, variables can be readily added or subtracted from the generative model as the literature evolves, and the effect of these changes on the probabilistic outcome (e.g., prognosis, illness trajectory, likely response to treatment) could then be assessed.

Phase Ia: Refinement and Testing Theories

In part because there is a solid foundation of evidence pointing to the likely role of both cortico-striato-thalamo-cortical circuits and anxiety/fear-related circuits in the pathophysiology of OCD, the field has likewise converged on several theories regarding the evolution of OCD symptoms. For purposes of this example, we will focus on three main theories which suggest that OCD results from (a) dysfunction in goal-directed behavior systems, (b) overactive/dysfunctional habit systems, and/or (c) dysregulation of threat processing and/or anxiety expression.

Recent work has suggested that OCD symptoms can result from an imbalance in the systems guiding action selection: the goal-directed "model-based" and the habitual "model-free" systems. Initial evidence has pointed mostly to dysfunction in the model-free, or habit, system (Gillan et al. 2011, 2014, 2015), with OCD patients being more prone to forming habits both in neutral conditions and "in avoidance" (i.e., to avoid a perceived threatening stimulus, such as a shock). This would suggest that excessive "model-free" action selection could be used as a latent construct reflecting a hidden physiologic state in our nosological framework of OCD. However, other evidence from both clinical deep-brain stimulation studies (Greenberg et al. 2006, 2010; Goodman et al. 2010; de Koning et al. 2011; Figee et al. 2014; Mantione et al. 2015), imaging studies (Baxter et al. 1988; Swedo et al. 1989; Rauch et al. 1994; Alptekin et al. 2001; Mataix-Cols et al. 2004), and preclinical studies in mice (Ahmari et al. 2013) suggests that dysfunction in medial orbitofrontal and ventral striatal regions linked to goal-directed systems can lead to abnormal compulsive behaviors. Thus it would be useful to be able to determine how changing the balance

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.

between the model-based and model-free systems would affect symptom presentation, diagnosis, and other observable entities. Using our model, this could be accomplished using different latent variables for model-based, model-free, or arbitrators between model-based and model-free systems (Dayan and Balleine 2002; Redish et al. 2008; Dezfouli and Balleine 2012; Wunderlich et al. 2012a; Dolan and Dayan 2013; Dayan and Berridge 2014; Lee et al. 2014).

Finally, there has been debate, in part heightened by the recent separation of OCD from other anxiety disorders in DSM-5, about whether abnormalities in anxiety regulation or threat processing play a pathologic role in OCD. A potential role for anxiety dysregulation in the pathogenesis of OCD is supported by factors including its prominence as a clinical symptom in OCD patients, observations of fear-conditioning abnormalities (Milad et al. 2013), and enhanced avoidance habits in OCD (Gillan et al. 2014), and reports of trauma-induced OCD (Dykshoorn 2014). Because the observable entities of symptoms (including anxiety), brain metabolic state (functional imaging), and neurocognitive task performance are all known in the recent studies of habit formation in OCD, these data could potentially be used to validate the efficacy of the model in one defined case.

Phases II and III: Application and Refinement

As described above, the potential power of this Bayesian framework lies in the fact that it is agnostic about the latent constructs and theories which make up a particular mental illness. It therefore easily incorporates the possibility that a disease currently classified as a single entity in DSM-5 could be broken down into several different theory-based categories with different disease and treatment trajectories. In fact, in an ideal case, the framework would actually help to identify valid patient subgroups through an iterative process of including progressively higher quality data in the model as the field evolves, and examining treatment outcomes in subpopulations of patients. Ultimately, delineating whether valid patient subcategories exist, or identifying which (if any) of the theories described in Phase Ia underlies pathophysiology for a particular patient, could be extremely important for guiding treatment. For example, a patient with excessive habits leading to an increased propensity to develop compulsions might benefit most from habit-reversal training, whereas a patient with high levels of anxiety might benefit more from treatment with selective serotonin reuptake inhibitors (SSRIs) and classic exposure therapy with response prevention.

There are, of course, potential limitations in using OCD as a test case. For example, first, unlike several other psychiatric disorders, there is limited data available on longitudinal trajectories in OCD throughout the life span, due to a lack of large-scale multisite longitudinal studies. Second, though similar brain regions are consistently highlighted across multiple imaging studies, behavioral studies of neurocognitive functions such as set shifting, response inhibition,

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.

and reversal learning are more variable. Although these factors may currently limit the richness of the data that can be incorporated into the model, they do not affect utility. Indeed, it is important to emphasize that "bad data" or data not relevant to the probable outcomes will automatically be removed from the integrative framework.

Special Considerations: Comorbidity

OCD is a particularly useful case example for exploring the issue of how our Bayesian framework can integrate comorbidities. OCD is highly comorbid with several other DSM-5 diagnoses, including major depressive disorder, panic disorder, Tourette syndrome, hoarding disorder, and obsessive-compulsive personality disorder (Murphy et al. 2013). In fact, hoarding disorder was only separated from OCD in the most recent edition of DSM based on findings from neuroimaging studies which identified distinct neurobiological substrates and differences in treatment response. This has raised the question of whether OCD patients with these comorbid disorders should be considered as belonging to separate diagnostic categories. This is an important consideration because there is some evidence that OCD patients with different comorbidities may have distinct responses to treatments. For example, it is commonly known that hoarding, which used to be considered an OCD symptom, is more resistant to both pharmacotherapeutic and psychotherapeutic treatment than classic OCD symptoms (Bloch et al. 2014; Mataix-Cols 2014). In addition, a recent study suggested that augmentation with atypical antipsychotic medications may be most useful in OCD patients who have comorbid tics (Bloch et al. 2006). Within the context of our proposed framework, comorbid conditions can be easily included in the generative model as further empirical observations (see Figure 10.6). By combining this with other observations, including neuroimaging data, genetic information, and neurocognitive task performance, we can determine the impact of comorbidities on OCD illness trajectories and potentially glean information about the most effective treatment interventions through the iterative process described above.

Discussion

The Bayesian Integrative Framework that we propose inverts the standard model of nosology. That is, rather than subscribing to the notion that particular entities that are classified by diagnostic categories cause symptoms, we are proposing that diagnostic classification and symptoms are a consequence of latent variables (or constructs) which themselves are caused by evolving but hidden pathological states. We consider the problem of nosology as modeling the diagnostic (and prognostic) process, where diagnosis is an observation or an outcome as opposed to a cause. Implicit in this framework is the mapping

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.

from the hidden or unobservable causes of psychopathology to observable symptoms and signs, and "good practice" diagnostic outcomes. This framework, therefore, accommodates risk assessment and the resolution of ambiguous nosological problems (e.g., comorbidity). Crucially, it accommodates, but is not limited to, currently accepted diagnostic categories (e.g., DSM). Further, it accounts for developmental and longitudinal trajectories of mental illness. Finally, it furnishes a formal link between putative constructs (e.g., RDoC constructs) and clinical categories, thereby harnessing the complementary perspectives afforded by dimensional and categorical approaches. Below we will further discuss the value of this model over the current diagnostic process and practical considerations for the implementation of this framework.

How Do We Define Success of a Diagnostic System?

Given that diagnostic systems such as nosologies exist primarily to assist clinicians in the management of their patients, the measure of the success of a diagnostic system can be framed as the degree to which it achieves that goal. There are several ways in which a diagnostic system can assist clinicians, all of which are covered under the broad rubric "clinical utility."

The first such use, the one in which categorical classifications like the DSM have been the most successful, is facilitating communication among clinicians, between clinicians and patients/families, and between clinicians and administrators. It will be important to retain this strength in any newly developed scheme.

A second use is to help clinicians select the optimal treatment for a patient. Ideally, making a diagnosis would be an initial critical step in guiding the choice of treatment (i.e., if the clinician makes a diagnosis of X, he or she can be confident that treatment Y is very likely to be effective). In actuality, however, there is an uncertain relationship between the current DSM diagnostic categories and treatment. Most psychiatric treatments are at least somewhat effective for a variety of categories that cut across the various DSM categories: SSRIs work for depressive disorders, OCD, PTSD, premenstrual dysphoric disorders, and others (Wagstaff et al. 2002; Saxena et al. 2007; Rapkin and Winer 2008). Moreover, the effectiveness of a particular intervention in treating a particular diagnosis has been disappointing for some patient subgroups within a diagnostic category. In treating major depressive disorder, the likelihood of any significant clinical benefit from antidepressant treatment is no better than 70%, and the likelihood of full remission is far lower (e.g., Rush et al. 2006b; Khin et al. 2011).

A third use is to help clinicians predict the future course and outcome of a psychiatric presentation (e.g., to inform the patient how likely it is for the symptoms to remit, or get worse, over time, as well as what environmental or other factors are likely to make the symptoms worse, or better). As with predicting treatment response, because of the range of potential course and outcome trajectories associated with each disorder, meeting criteria for a DSM

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.

category provides limited information in terms of predicting a future course. For both treatment and prognostic prediction, much of the problem stems from the fact that the diagnostic categories are essentially "black boxes" which obscure crucial mechanistic differences between individual patients included in a particular category. This suggests that in order for a diagnostic system to be successful, it must incorporate mechanistic processes in the diagnostic formulation, which is exactly what the proposed framework will do.

One example of a framework that incorporates the kind of categorical diagnostic decisions that are essential to clinical practice, while grounding these decisions in a biologically meaningful process, is the "harmful dysfunction" model (Wakefield 1992a, b, 2007). The harmful dysfunction model defines psychiatric disorders as requiring both a value judgment (harm, negative consequences) and a dysfunction judgment (when the condition represents dysfunction of a naturally selected mechanism or trait). Importantly, harm is not a scientific question but rather a value question. Harm is defined as something that causes distress or is socially disvalued, thus "harm" inevitably involves a value judgment. The harmful dysfunction analysis specifies that neither harm without dysfunction (e.g., procrastination, illiteracy, grief) nor dysfunction without harm (e.g., synesthesia) is a disorder. While the harmful dysfunction analysis provides an illustrative example of incorporating mechanisms into categorical diagnoses, this model has not been widely accepted due to the limited knowledge about the nature of "natural functions." Regardless, nosologies that reflect the harmful dysfunction combinatorial approach possess major advantages. For progress to occur, such nosologies should utilize the knowledge we have about the brain to inform diagnostic decisions, as we have proposed to do with the Bayesian Integrative Framework.

Problems with the Current Diagnostic Process

Creating Categories out of Continuous Data

Research demonstrates the continuous nature of many psychiatric problems, without a natural breakpoint between health and disease (Fergusson and Horwood 1995; Kendler and Gardner 1998; Fergusson et al. 2005). As a result, a major problem faced by diagnosticians arises from the need to impose a categorical structure on information that is largely continuous in nature (Van Os et al. 1999). For example, when we examine overall levels of multiple symptoms related to anxiety, we find that we can arrange individuals in a continuous distribution, from few to many symptoms. Moreover, when we examine the predictive relationship between symptom number and various external validators, such as outcome or treatment response, we see no natural break point in these relationships. More symptoms predict worse outcome or treatment response in a continuous fashion.

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.

Another example of continuous data that can be problematic relates to the risk of a false-positive or false-negative diagnosis. When an individual receives a psychiatric diagnosis in error, this carries risk, both due to delivery of treatment and its resulting harm as well as to the stigma associated with many psychiatric diagnoses. Again, this gives a set of continuous relationships, such that the more extreme the treatment, the greater the risk. Other areas of medicine provide guidance for the clinician in the circumstances that currently confront the psychiatrists. Clinicians can derive categories based on the point on a continuous scale where the benefits associated with treatment outweigh the risks of false positives. Such an approach has been, for example, used in obstetrics. In some consensus guidelines, age 35 had been considered a breakpoint when considering the appropriate age for amniocentesis. This age was a point where the rate of a positive diagnosis of a genetic anomaly became greater than the rate of miscarriage associated with the procedure. However, this crossing of risks assumes that the parent's judgment of negative consequences is equivalent between genetic anomaly and miscarriage. Clearly, in current practice, other factors influence decisions about amniocentesis. Discussing the risks separately allows the parents to make decisions based on their own value judgments of risk. Similar computations apply for the treatment of hypertension, where treatment is initiated when the benefits, in terms of reducing risk, outweigh the risks associated with side effects.

Psychiatry currently faces a few problems in the applications of this approach. Importantly, we need more data to precisely quantify the nature of risks associated with various levels of symptoms, various treatments, and false positive diagnoses. In addition, we need a process for identifying thresholds at which risks from declaring a positive diagnosis outweigh the risks from a false positive diagnosis. The proposed Bayesian Integrative Framework addresses these problems and provides guidance for better predicting outcomes and treatments in a way that maintains the continuous nature of risk and thus minimizes the adverse consequences of false-positive diagnoses.

Accounting for Trajectories

Most psychiatric disorders have a longitudinal evolution, with underlying risk factors and symptoms changing over time. Still, current diagnostic systems incompletely incorporate a developmental perspective and do not fully utilize repeated, longitudinal observations. This is important because clinical decisions are often based on estimating the probability of future trajectories for a patient that can only be determined by repeated observations. For example, the first episode of a mild disorder in adolescence (e.g., subthreshold depression) might spontaneously remit, progress to a full major depressive episode, or herald future bipolar disorder (Rutter et al. 2006). Many psychiatric disorders are preceded by earlier difficulties in childhood (Kim-Cohen et al. 2003; Pine and Fox 2015). For example, mood disorders are commonly preceded by anxiety

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.

and behavioral problems (conduct disorder). Schizophrenia can be preceded by earlier childhood symptoms and neurodevelopmental impairments involving anxiety, inattention, as well as cognitive, motor, language, and social communication impairments (Kim-Cohen et al. 2003; Rutter et al. 2006; Dickson et al. 2012). Most individuals with childhood anxiety or conduct disorder, however, do not go on to develop serious forms of mood disorder (e.g., bipolar disorder), and most children with neurodevelopmental impairments do not develop schizophrenia.

Importantly, these variable trajectories pose problems for clinical decision making. Clinical symptoms, such as depression, observed at a particular point in development, can emerge through multiple earlier developmental trajectories—a process often referred to as *equifinality* (see Cicchetti and Rogosch 1996). The idea of equifinality can be applied to multiple factors beyond development. Thus, risk factors or brain dysfunction can also be viewed from this perspective. Two distinct risk factors or two different types of brain dysfunction can predict the same symptomatic presentation. Heterogeneity from equifinality arises when different pathophysiological processes give rise to the presentation of similar symptoms. Another form of heterogeneity arises from a process termed *multifinality*. This means that one risk factor, such as a traumatic life event, can give rise to multiple different outcomes. Like equifinality, the idea of multifinality can be applied to development, whereby one developmental profile has many different outcomes (as discussed above). This can also be applied to risk factors (e.g., genetic variants) and brain function, whereby a single risk factor or a single type of brain dysfunction gives rise to many different outcomes.

The proposed integrative framework will provide a means to incorporate and appropriately weight all of the available observations (e.g., clinical variables, cognitive variables, family history of bipolar disorder), and, in turn, inform clinicians as to the probability of the trajectory their patient is most likely to follow. This will be important in influencing decisions about whether or not to intervene, balancing the risks versus side effects of diagnostic labeling or treatment, choosing an appropriate intervention, and selecting the intensity of treatment and follow up—ultimately improving diagnostic and treatment outcomes.

Knowing if the Model Is Right or Wrong

We have cast the problem of optimizing a Bayesian Integrative Framework among generative models of diagnostic/prognostic outcomes and their associated symptoms and signs. So how do we know if a model generated from this framework is correct or incorrect? Strictly speaking we can dissolve this question by noting that all models are wrong but some have much more evidence than others (recall that model evidence scores the goodness of a model in terms of the right balance between the accuracy of fitting some data and the complexity of the model; see above).

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.

The notion of right and wrong presupposes that there are only two models—a correct and an incorrect model. Formally, the idea of right and wrong models is implicitly assumed in classical statistics and corresponds to the null and alternative hypotheses, respectively. We reject the null model as wrong if there is enough evidence to make one chosen statistic sufficiently large. However, in Bayesian model comparison the number of competing models or hypotheses can be much greater than two, and each model has its own evidence. Bayesian model comparison then reduces to selecting those models that have the greatest evidence. This is a simple thing to do because the difference in log evidence between one model and another corresponds to the log of their relative (marginal) likelihood. For example, if the best model has a log evidence of three or more, relative to the next best model, then the best model is twenty ($\exp(3) = 20$) times more likely than all of its competitors. This, however, is a relative statement; it only pertains to the set or space of models considered. In this sense, there is no right or wrong model—only models which are better or worse at explaining any given data in a parsimonious fashion. In this regard, the current integrative framework will allow us to identify the "best" fit model on a case by case basis and will, undoubtedly, provide a foundation to improve upon the current diagnostic system in psychiatry.

Potential Problems

The implementation of this framework faces technical and community acceptance challenges—challenges that will need to be addressed with outreach, training, and continuous communication between the fields of computation, psychiatry, and neuroscience. A number of advances have already developed in this regard, with the emergence of new training programs and funding opportunities and new journals (e.g., *Computational Psychiatry*). In addition, a Transcontinental Computational Psychiatry Workgroup has recently emerged out of this Ernst Strüngmann Forum. This group consists of scientists from the fields of computation, neuroscience, and psychiatry who have begun to convene on a regular basis to discuss and advance the field of computational psychiatry. In turn, new conferences are being developed and the field is gaining increasing recognition. Moreover, we are aware that there will be financial and political barriers that will need to be overcome in adopting this framework. Finally, it is important to note (as outlined in the discussion of the *Phases* above) that multiple iterations of this model will be required for continual updates and improvements based on the information that we have at any given time.

Summary

In this chapter we have proposed and described the Bayesian Integrative Framework—a novel integrative framework intended to integrate neuroscience

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.

and clinical psychiatry research, enhance the current diagnostic process, and improve treatment outcomes in psychiatry. We have identified several important features of the proposed framework that will allow us to circumvent some of the challenges currently being faced in the field of psychiatry. For one, we will be able to integrate mechanistic processes in diagnostic formulation; that is, we can incorporate any knowledge we have about underlying pathophysiology to inform diagnostic decisions. Indeed, the Bayesian Integrative Framework provides the necessary bridge between putative constructs (e.g., RDoC) and clinical diagnoses, thereby linking the complementary perspectives afforded by dimensional and categorical approaches. It also allows us to incorporate many different flavors of data at multiple layers. One of the most valuable features of this framework is perhaps its ability to account for and incorporate longitudinal trajectories that may nuance diagnosis, prognosis, and treatment. This framework will yield a better understanding of individual differences and, importantly, how individual differences in brain function give rise to individual differences in behavior.

Acknowledgments

We would like to thank David Redish and Joshua Gordon for their foresight in proposing this Forum and their dedication to making it happen. We also thank the other Program Advisory Committee members, especially Julia Lupp, for organizing and hosting the Forum. We would also like to thank the staff for their administrative help with this chapter, particularly Eleanor Stephens. Finally, we would like to thank all of the Forum participants for helpful discussions and comments on early drafts of this chapter.

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.