ANNUAL REVIEWS

Annual Review of Neuroscience

Smartphones and the Neuroscience of Mental Health

Claire M. Gillan[1] and Robb B. Rutledge[2,3,4]

[1] School of Psychology, Trinity College Institute of Neuroscience, and Global Brain Health Institute, Trinity College Dublin, Dublin 2, Ireland; email: gillancl@tcd.ie

[2] Departments of Psychology and Psychiatry, Yale University, New Haven, Connecticut 06520, USA; email: robb.rutledge@yale.edu

[3] Max Planck UCL Centre for Computational Psychiatry and Ageing Research, University College London, London WC1B 5EH, United Kingdom

[4] Wellcome Centre for Human Neuroimaging, University College London, London WC1N 3AR, United Kingdom

Keywords

smartphone, psychiatry, mental health, cognitive neuroscience, big data, longitudinal

Abstract

Improvements in understanding the neurobiological basis of mental illness have unfortunately not translated into major advances in treatment. At this point, it is clear that psychiatric disorders are exceedingly complex and that, in order to account for and leverage this complexity, we need to collect longitudinal datasets from much larger and more diverse samples than is practical using traditional methods. We discuss how smartphone-based research methods have the potential to dramatically advance our understanding of the neuroscience of mental health. This, we expect, will take the form of complementing lab-based hard neuroscience research with dense sampling of cognitive tests, clinical questionnaires, passive data from smartphone sensors, and experience-sampling data as people go about their daily lives. Theory- and data-driven approaches can help make sense of these rich data sets, and the combination of computational tools and the big data that smartphones make possible has great potential value for researchers wishing to understand how aspects of brain function give rise to, or emerge from, states of mental health and illness.

Review in Advance first posted on February 8, 2021. (Changes may still occur before final publication.)

Contents

INTRODUCTION

As our understanding of the neurobiological and cognitive correlates of mental health and mental illness has grown through decades of research, one thing has become clear: Things are more complicated than we might have hoped. The notion of one-to-one mappings between abnormalities in specific brain areas or cognitive markers and individual categories from the *Diagnostic and Statistical Manual of Mental Disorders* (DSM-5) (APA 2013) has all but been abandoned. Much like the field of neuroscience overall (Button et al. 2013), clinical neuroscience research has been substantially underpowered (Marek et al. 2020), and the findings from many studies do not hold up when subjected to large-scale replication attempts (Rutledge et al. 2017) or meta-analyses (Müller et al. 2017, Widge et al. 2019), and when they do, effect sizes are small (Marek et al. 2020) and dependent on disorder versus healthy control comparisons (Davidson & Heinrichs 2003, Hoogman et al. 2019), rather than being specific to one diagnostic category over the next (Bickel et al. 2012, Gillan et al. 2017, Lipszyc & Schachar 2010). Excitement about the role of candidate genes or gene–environment interactions in major psychiatric disorders has been replaced with the acknowledgment that complex mental health conditions are massively polygenic, and single genes likely carry very small individual risk (Farrell et al. 2015, Flint & Kendler 2014). Likewise, in terms of environmental influences, childhood adversity and stress confer similar generalized risk for psychopathology (Kessler et al. 1997), and while substantial, it is far from deterministic. These influences may compete and interact with myriad other factors such as diet (O'Neil et al. 2014), exercise (Chekroud et al. 2018), the gut microbiome (Kelly et al. 2015), social isolation (Richard et al. 2017), urban living (Paykel et al. 2000), socioeconomic status (Lorant et al. 2003), sleep (Ford & Kamerow 1989), drug use (Jané-Llopis & Matytsina 2006), alcohol (Weitzman 2004), and cigarettes (Fluharty et al. 2017). For many of these factors, it remains challenging to arbitrate between causation versus selection-based explanations (Goldman 1994) and to develop strong causal models. The role that cognition plays in mental illness is even less-well understood because large epidemiological-style studies of brain process are challenging to conduct.

Here we argue that in order to develop robust neurocognitive models of mental illness, we must invest in new methods that can deliver on substantially richer, multivariate data sets and larger samples than are feasible in the traditional small, single-site studies that dominate the field (**Figure 1**).
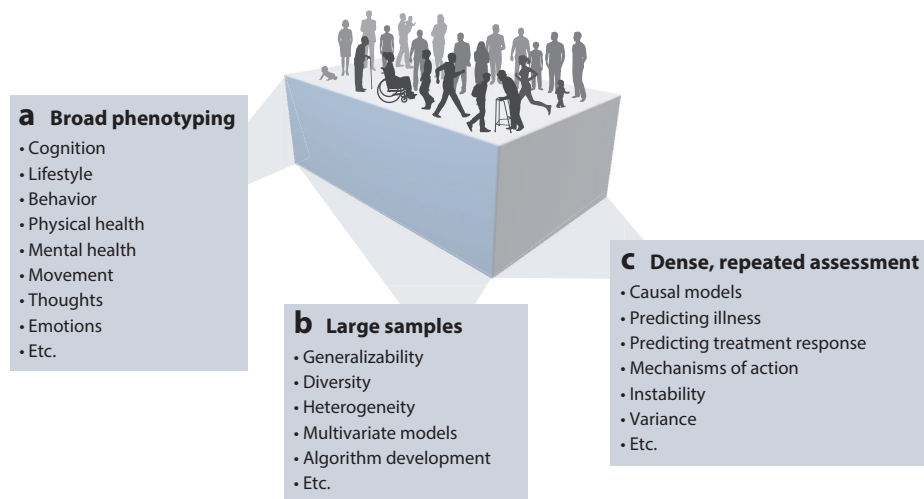
Figure 1

Smartphones can deliver high-dimensional data sets relevant to the neuroscience of mental health and illness. (*a*) One of the key benefits of developing a smartphone-based infrastructure for neuroscience research in mental health is that it can increase the breadth of data gathered on any single individual (broad phenotyping), including passive data (e.g., movements in the natural world, social connections) and dense self-report (e.g., daily fluctuations in mood). This, we believe, will greatly enhance research conducted at multiple levels of analysis, including research in smaller samples that also have harder neuroscientific measurements (e.g., structural and resting-state neuroimaging). (*b*) By increasing the overall sample size (large samples), we can test the extent to which our models generalize, account for diversity in populations, can be related to heterogeneity in symptoms, and can support multivariate analyses and algorithm development needed to bring together complex, interactive data sets. (*c*) Finally, smartphones allow us to enhance the depth of the assessments we gather (dense, repeated assessment). This is crucial for a field seeking to move beyond cross-sectional methods toward longitudinal methods that can help us to understand causation, make predictions, and delineate mechanisms of change over multiple months.

New approaches must be capable of capturing numerous interacting and confounding variables within the same individual, and crucially, they must facilitate following large cohorts through time. The vast majority of research in psychiatric mechanisms is cross-sectional—this, we believe, presents the most significant barrier to the delivery of neuroscience-informed clinical tools. If we hope to translate cognitive insights into clinical treatment, a major paradigm shift is needed that can move the field beyond the descriptive and toward the predictive (Browning et al. 2020).

To this end, we focus on smartphones as a new methodology for basic research in neuropsychiatry that can dramatically increase the depth and breadth of research and encourage a shift away from cross-sectional research to longitudinal designs that are essential for clinical translation. In doing so, we must, for practical reasons, feature cognitive, behavioral, and clinical measurements more than others. Later, we discuss how these data can be linked to the harder tools of neuroscience but emphasize that this is not the only, or indeed primary, goal of this endeavor. Much of the benefit/opportunity of a smartphone-based approach to clinical neuroscience is to enrich research within this higher-order level of analysis because (*a*) much is to be gained for neuroscience in the study of behavior alone and (*b*) for those most interested in more direct measures of brain function, it is nonetheless these levels of analysis that neuroscientists seek to explain, understand, and/or predict.

Although the uptake of smartphones for research in this field is relatively new, we draw on several recent examples that highlight the potential of this new discipline. Because research in this area is in its infancy, we also discuss internet-based research methods more generally, which have risen in popularity for psychiatry research in the last five years (Gillan & Daw 2016) and that share some of the advantages of smartphone-based approaches, particularly for testing large and diverse samples. Crowdsourcing platforms that support browser-based testing such as Amazon Mechanical Turk (AMT) and Prolific may soon comprise the majority of all cognitive neuroscience studies (Stewart et al. 2017). We lay out a roadmap for a natural extension of this methodology with substantial added value: cognitive neuropsychiatry research in the age of smartphones.

DATA QUALITY

Can We Collect High-Quality Data Remotely?

Several validation studies have demonstrated that internet-based, remote cognitive testing yields reliable and valid data, whether collected via crowdsourcing platforms (Crump et al. 2013, Goodman et al. 2013) or more general browser-based methods (Casler et al. 2013, Germine et al. 2012). Such data are often thought to be noisier than in-person data, but there is evidence that this is to some extent task dependent (Crump et al. 2013). For example, the association between normal variation in compulsivity and deficits in a relatively complex cognitive capacity [model-based planning (Daw et al. 2011)] requires 461 students in person to have 80% power to detect an effect at $p < 0.05$ (Seow et al. 2020), while online via AMT, the required N rises approximately 30% to 670 (Gillan et al. 2016). In a recent developmental study, it was shown that while 15 subjects per group are required to observe a change in model-based planning from childhood to adolescence, 21 are required on AMT (Nussenbaum et al. 2020). These differences are thus consistent, but relatively modest, and could be due in part to differences in subject motivation or other demographic differences between university-based versus AMT samples.

Cognitive task data gathered via smartphone are likely to be even noisier because the testing environment is less controlled as subjects participate on the go. Furthermore, unpaid participants may be more likely to quit tasks that they do not enjoy, leading to incomplete data sets. Over 40,000 people downloaded The Great Brain Experiment in the first month after release, and approximately 20,000 of those submitted complete data for at least one 5-min task. Smartphone data showed similar effects to those observed in the lab and with comparable quality to in-person studies across multiple domains of cognition (Brown et al. 2014). For example, the effect of distraction on working memory performance was similar in magnitude in over 3,000 people assessed via smartphone (Cohen's d = 0.42) and in 21 participants tested in the lab (Cohen's d = 0.37) (Brown et al. 2014). Out-of-sample model predictions for mood dynamics during a risk-taking task were higher, but not substantially so, in two lab samples (mean model fit r^2 of 0.29 and 0.33) compared to a much larger and more diverse smartphone sample (mean model fit r^2 of 0.24) (Rutledge et al. 2014). In contrast, effects sizes for the stop-signal reaction time (SSRT) task were three times larger for in-person compared to smartphone samples (Brown et al. 2014). A major strength of smartphone-based testing is that increased noise can be mitigated by collecting substantially larger samples than are feasible to collect in person because the cost of testing additional subjects can be negligible. For example, for the observed reduction in the key effect size associated with SSRT, adequate power requires one order of magnitude more participants, but the sample size collected via smartphone was approximately 10,000 (Brown et al. 2014), more than two orders of magnitude larger than typical in-person cognitive studies.

Aside from sample size gains, smartphone-based studies provide the opportunity to easily evaluate the robustness of links between symptoms and task performance with multiple task variations. A/B testing also allows researchers to refine the experimental design (Daniel-Watanabe et al. 2020) and improve upon critical reliability metrics essential for between-subject designs (Hedge et al. 2018). Though not outside the scope of lab-based experiments, smartphone tasks that are gamified naturally lend themselves to being adaptive to a user's performance. For example, using Bayesian adaptive algorithms to present maximally informative options for the estimation of decision model parameters can greatly increase the efficiency of data collection (Pooseh et al. 2018).

What About Clinical Data?

The extension of this methodology to psychiatry research is rising, but perhaps more slowly than for cognitive science research (Chandler & Shapiro 2016). This slow pace may be because clinical researchers harbor doubts that such an approach is valid because remote formal diagnosis may be impossible. While the self-report measures that are most easily collected are fundamentally different, ample arguments support a move toward greater use of self-report measurements in psychiatry research. Without digressing into the broader critiques of the validity of DSM-5 constructs (Fried & Nesse 2015, Haslam et al. 2012, Kapur et al. 2012), it is important to highlight some salient issues related to reliability. In the DSM-5 field trials, which saw two clinicians perform separate diagnostic interviews with the same patient [interval ranging from 4 h to 2 weeks (Clarke et al. 2013)], the inter-rater reliability of clinician-assigned disorders was low for some of the most prevalent and most-studied disorders, including major depressive disorder and generalized anxiety disorder (GAD) (Regier et al. 2013). Without this basic psychometric property, studies aiming to link brain changes to disorder categories can never show strong associations. In contrast, self-report assessments of the same constructs can perform considerably better, whether collected in person or online. For example, studies run on AMT find high 1-week test-retest reliability for the Beck Depression Inventory ($r = 0.87$) (Shapiro et al. 2013) and 3-week test-retest reliability for the Big Five ($r = 0.85$) (Buhrmester et al. 2011). Self-report questionnaires have the advantage of avoiding variability across clinician (in interpretation of patients' responses to interview probes) and within clinician (the reliability of that interpretation over time).

Concerns about the utility and reliability of DSM-5 categories, particularly for the most common disorders like anxiety and depression, suggest that using self-report assessments rather than clinician-assigned diagnoses could actually improve our ability to relate changes in brain function to specific aspects of psychopathology. These assessments have the distinct advantage that they can easily collect remotely with much less effort than structured clinical interviews. A recent study found empirical support for this possibility (Gillan et al. 2019). A structured telephone-based diagnostic interview established DSM-5 diagnoses of either GAD, obsessive-compulsive disorder (OCD), or a combination of the two in a sample of 285 patients. Subjects completed multiple self-report clinical questionnaires and an online cognitive test of model-based planning (Daw et al. 2011), which had been previously linked to compulsive disorders in a series of case-control studies (Voon et al. 2014). No difference in cognitive performance was observed between these patient groups. However, higher self-report levels of compulsivity across the entire sample (collapsed across diagnostic categories) were associated with reductions in model-based planning ability. These data suggest that self-report clinical data might provide a closer mapping to underlying brain changes than the diagnosis one is assigned.

Although self-report has advantages for reliability, there are cases where we might expect self-report responses to be less valid than clinician-assigned diagnosis, and this warrants more careful study. One example is the study of mental health issues characterized by a lack of insight; for

example, Shapiro et al. (2013) noted implausibly high rates of mania in an online sample. Another is when eligibility checks encourage deception, for example, by indicating to participants that they must endorse certain symptoms at prescreening in order to participate, one increases the chances of malingering (Chandler & Paolacci 2017). Attention to these and other potential issues will be critical as we develop new testing protocols and study designs specific to smartphone-based research. Of course, in-person research can also be susceptible to experimenter demand effects, and careful consideration of possible effects is essential for both in-person and remote studies. Beyond data integrity, online research methods present important ethical dilemmas that are the subject of current debate. Most prominent among these is the issue of clinical responsibility over subjects who may be anonymous participants or in locations far removed from the researcher. This complex issue is beyond the scope of this paper but likely to gain prominence as online protocols for research in this area become commonplace. When it comes to the data themselves, however, we can say with confidence that self-report clinical data collected in online studies have been shown to be valid and reliable (Chandler & Shapiro 2016), providing a compelling justification for the proliferation of smartphone (and other internet-based) studies of psychiatry and cognition.

Are Smartphone Samples Representative?

An important but oft overlooked requirement for clinical translation is that the findings from our studies generalize to new settings and, most importantly, new samples. Most in-person cognitive studies evaluate primarily young and educated samples (Henrich et al. 2010). Online samples are typically more diverse (Buhrmester et al. 2011), and therefore the results from smartphone studies may be more likely to generalize to clinical samples. Moreover, smartphone studies facilitate comparison across geographic regions with relative ease. This fact will be crucial as we think about how findings from primarily Western samples translate to developing countries, where, for example, the incidence of dementia is expected to rise most over the next 50 years (Kalaria et al. 2008).

Though age is considered a barrier to online research, with limited participation expected from elderly users, older adults increasingly use smartphones and have interest and time to participate in research, particularly where the topic is of relevance. For example, of the first 4,000 registered users of the Neureka app, which has a partial focus on early dementia detection, the average age was 39 and 25% were 50 years or older (**Figure 2a**). These distributions are much more representative than lab-based or even AMT samples (**Figure 2a**). Testing of thousands of elderly participants via The Great Brain Experiment has shown that ageing leads to working memory being increasingly compromised by distractors presented during encoding (McNab et al. 2015) and that ageing reduces risk-taking for rewards (Rutledge et al. 2016).

In addition to achieving a greater spread of ages, it is perhaps also notable that users of the Neureka app have a broader distribution of depression symptomatology ($N = 1,500$) than in-person or AMT samples (**Figure 2b**), which might be a feature of citizen science research, where subjects are usually not paid for participation. Marketing is likely more effective for individuals with a personal experience of mental health, potentially leading to oversampling of individuals with current depression. Smartphone-based assessment can also resolve potential confounds or provide important qualifiers to gender-based findings. A recent example was the finding of gender differences in spatial navigation ability in over 500,000 users of the Sea Hero Quest app (**Figure 3**). While this finding is relatively well documented in smaller face-to-face studies (Linn & Petersen 1985), Coughlan et al. (2019) used their large data set to reveal an important qualification—the difference between genders was partially explained by the extent of the gender inequality in the country from which data were drawn (Coutrot et al. 2018).
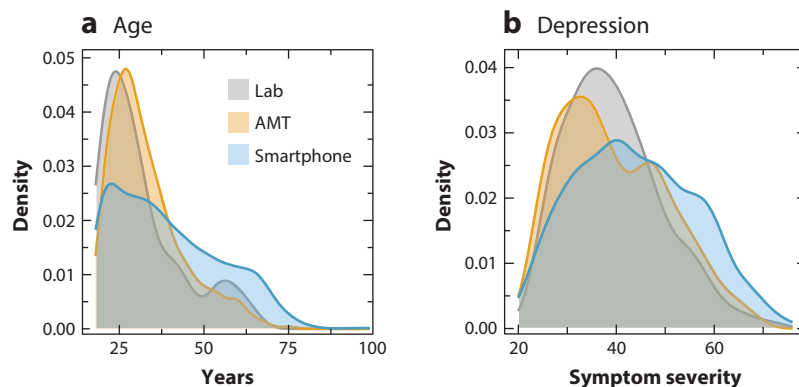
Figure 2

Demographic and clinical comparison of laboratory, online worker platform [Amazon Mechanical Turk (AMT)], and smartphone-based samples. (*a*) Density plot depicting the age profile of participants recruited in the laboratory ($N = 185$), participants recruited via AMT ($N = 1,413$), and registered users of the smartphone app Neureka ($N = 4,000$). The Neureka app achieved a much broader spread of ages in its first 4,000 early adopters compared with in-person and AMT samples. (*b*) Density plot of depression scores collected using the Zung Depression Inventory (Zung 1965) for subjects recruited in the laboratory ($N = 185$), through an online worker platform (AMT, $N = 1,413$), and using Neureka on their smartphones ($N = 1,500$). Depression scores have a broader distribution in the smartphone-based sample. The broader spread of demographics and symptoms in smartphone-based samples increases both statistical power and generalizability.

HOW SMARTPHONE DATA CAN ENHANCE COGNITIVE NEUROSCIENCE RESEARCH IN PSYCHIATRY

Repeated Within-Subject Assessment

A relative dearth of within-subject longitudinal assessments of cognition, behavior, and thought in contemporary neuroscience research represents a significant gap in knowledge that impedes our ability to develop and test causal models. Smartphones make possible dense experience sampling and the sort of repeated within-subject measurement that are essential for developing explanatory accounts of how changes in specific aspects of brain function might lead to, or result from, mental illness. Smartphone-based tools used to collect follow-up longitudinal data in participants in neuroscience experiments could be an inexpensive add-on to a neuroscience study that permits testing for links between brain activity and changes in symptoms over time. The growing utilization of personal smartphones makes possible large-scale inexpensive research in neuropsychiatry; cognitive tests and self-report measurements can be rolled out to thousands of research participants simultaneously and through time. To reduce the burden on participants, smartphones uniquely support the seamless integration of passive proxies for self-reported data points, and cognitive tasks can be designed to maximize engagement and enjoyment. Moreover, users can be prompted to participate as they go about their daily lives, increasing our ability to study the brain and behavior in naturalistic settings. As outlined in the Introduction, we largely focus our discussion on indirect measures of brain function (i.e., cognition), experiences (e.g., behavior), and self-report (e.g., emotions) that can be easily gathered via smartphone. Later, we describe how these data can be further enriched with more direct, hard tools of neuroscience, first in humans and ultimately across species.

The added value of rich within-subject assessment via smartphone is nicely illustrated in the findings from the Track Your Happiness project, which showed in 2,250 participants that mind
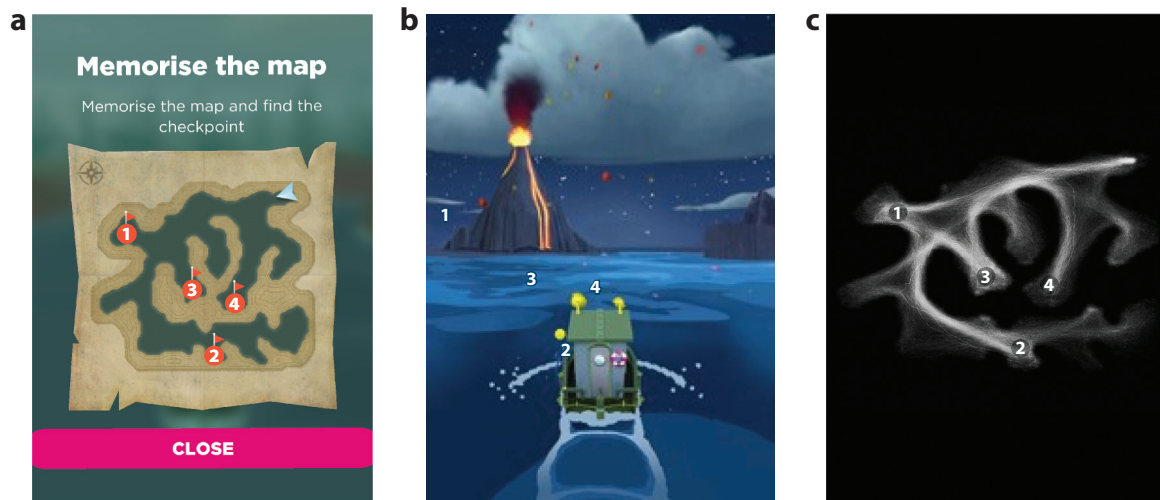
Figure 3

A smartphone-based spatial navigation task from Sea Hero Quest. (*a*) In the wayfinding task, participants are shown a map of checkpoints located in a gamified water maze. (*b*) The map then disappears, and they must navigate to those locations from memory. Success depends on multiple complex skills, including accurate interpretation of the map, multi-step planning, memory of the checkpoint locations and layout of maze, continuous monitoring and updating, and the transformation of a bird's-eye map perspective to an egocentric view as one steers the ship. (*c*) In all, 2,512,123 users played the game, and data were analyzed from 558,143 subjects with a sufficient number of levels completed. Here, randomly sampled data from 1,000 individual trajectories are superimposed. Overall, spatial navigation performance was quantified for wayfinding and a related task in the app, and the authors found that performance declined with age and was better in males relative to females. However, they found that this gender difference was smallest in countries with the greater gender equality (Coutrot et al. 2018). Figure adapted with permission from Coutrot et al. (2018).

wandering is associated with reduced happiness (Killingsworth & Gilbert 2010). While prior work had demonstrated this cross-sectionally (Smallwood et al. 2009), the availability of time-series data gathered through the app, within-subject, allowed the authors to conduct time-lag analysis, which revealed evidence for a perhaps unexpected direction of influence—mind-wandering precedes bouts of unhappiness. Using a similar experience-sampling approach, Villano et al. (2020) gathered dense samples of mood via smartphone in a student sample on days following their first viewing of a new exam grade. They then applied a popular construct in computational neuroscience, prediction error (when experiences differ from expectations), to this real-world situation. Outside of the typical lab setting, they found that prediction errors are critically important in dictating the mood of students. Specifically, Villano et al. observed that emotions following the issuance of grades depended more on prediction errors (i.e., whether their grade was better/worse than expected) than on the actual grade itself. This finding has proved generalizable to a variety of settings, for example, within the more constrained setting of a single game played on a smartphone. The Great Brain Experiment app asked individuals to repeatedly choose between safe and risky options, with the potential rewards and losses systematically varied to accommodate a large range of economic preferences (**Figure 4*a,b***). Chosen risky options had equal probabilities of either outcome (50/50) and were represented by spinners where the gambles were resolved after a brief delay. Participants started the game with 500 points, and the outcome of every trial counted toward the final score. Participants responded to the question "How happy are you right now?" at regular intervals by moving a cursor to indicate current affective state (**Figure 4*c,d***). Computational modeling showed that self-reported affective state depended on both expectations and prediction errors from the past 4–7 trials. Consistent with the real-world findings of Villano et al. (2020),
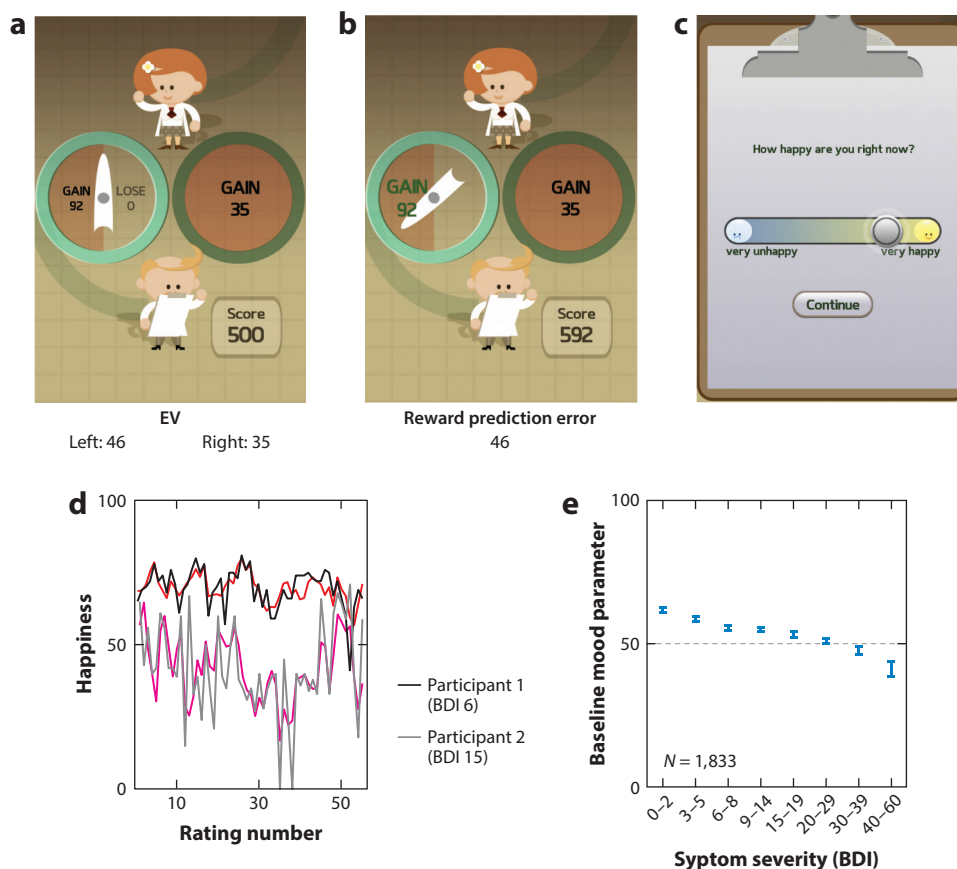
Figure 4

A smartphone-based cognitive task from The Great Brain Experiment. Participants chose between risky and safe options in a game in The Great Brain Experiment app (over 130,000 downloads) that varied in their expected value (EV). (*a*) The gamble on the left has a 50% chance of returning 92 points but a 50% chance of returning nothing. Although riskier, this corresponds to an expected value of 46 points, which is larger than the expected value of the option on the right (100% chance of 35 points). (*b*) Following their choice, they see the outcome of the gamble and experience a reward prediction error, which is the deviation between what was expected (EV = 46) and what they got (actual outcome = 92). (*c*) Subjects rated how happy they were after every few trials. (*d*) A computational model (*red lines*) was used to predict momentary happiness based on trial events, including the reward prediction error. Here we show its performance in two example participants, one with minimal depressive symptoms [Beck Depression Inventory (BDI = 6)] and one with significant depressive symptoms (BDI = 15). (*e*) The baseline mood of subjects during the game (estimated from the computational model) was correlated with depressive symptoms in both online and in-person (not pictured) samples (Rutledge et al. 2017).

momentary fluctuations in mood were shown to be dependent on prediction errors experienced during the game (Rutledge et al. 2014). Interestingly, while sensitivity to prediction errors was not linked to depression, baseline mood during the game was found to relate to depression symptom severity both in the lab and using smartphones (Rutledge et al. 2017) (**Figure 4e**). Modeling mood dynamics during reinforcement learning shows that baseline mood relates to depressive symptoms in volatile but not stable environments (Blain & Rutledge 2020). These studies illustrate how complementing between-subject assessments with more granular within-subject probes

provides a more nuanced view of how mood and reward sensitivity interact on different timescales and levels of abstraction.

The benefit of this sort of longitudinal data is further underscored by findings from the 58 Seconds app. Here, researchers tracked the sort of activities that over 28,000 users chose to engage in over an approximately 1-month period while in a good versus bad mood. Researchers were able to use these time-series data to ascertain that people actively select mood-increasing activities while in a bad mood (Taquet et al. 2016). Later, they showed that this mood homeostasis effect (choosing to engage in mood-modifying activities to stabilize mood) was reduced in people with overall low mood (in this same data set) and in people with a history of depression (in an independent data set) (Taquet et al. 2020). Together, these studies highlight how cross-sectional observations can be enriched from within-subject insights, offering new insights into underlying mechanisms. Numerous other examples of this have emerged, for example, the Mappiness app used geolocation tracking from over 20,000 users to show that happiness was higher when people were in natural compared to urban environments, controlling for weather, activity, companionship, and time of day (MacKerron & Mourato 2013). Another experience-sampling study used within-subject sampling to provide a window into the causal relationships that exist between sleep and mood. They found that day-to-day effects of sleep on mood are actually larger than the effect of mood on sleep (Triantafillou et al. 2019).

In some cases, within-subject analyses are important to shore up equivocal between-subject effects. For example, a cross-sectional analysis of nearly 27,000 US and UK users of The Great Brain Experiment app showed that risk-taking in trials with potential losses increased with time of day (Bedder et al. 2020). Computational modelling using prospect theory suggested that this pattern could be explained by a decrease in loss sensitivity that occurs as the day wears on, making large potential losses less aversive. Although an interesting possibility, with cross-sectional data only, it is difficult to know whether this is truly an effect of the time of day or rather reflects between-subject differences in the sorts of individuals who prefer to play at day versus at night. Importantly, a within-subject analysis of 2,646 users playing twice on different days between 8 a.m. and 10 p.m. identified a similar effect, increasing confidence that the time-of-day findings were not due to differences in diurnal patterns of users.

Neuromodulators are known to fluctuate throughout the day (Fuller et al. 2006), but the consequences for human behavior are unclear. A computational link between serotonin and loss-related behaviors (Dayan & Huys 2008) provides a potential explanation: higher tonic serotonin late in the day could reduce the dynamic range of serotonin transmission and reduce the aversiveness of potential losses. Neuroscientific data will be needed to confirm these predictions. Recent identification of profound reductions in global signal amplitude in resting-state brain network connectivity profiles over the day (Orban et al. 2020) suggests that these findings may be only the first of many discoveries linking circadian rhythms to human behavior and symptoms once adequate large-scale samples are acquired via smartphone.

Smartphones can also be used to repeatedly assess cognition in relation to ongoing treatment with pharmacological agents, a major gain for basic neuroscientific as well as clinical research. Selective serotonin reuptake inhibitor (SSRI) antidepressant drugs can take 4–6 weeks to take their effect, and we know surprisingly little about how clinical benefit is achieved. One popular theory is that these drugs positively bias one's response to affective information, having the effect of gradually improving mood (Pringle et al. 2011). Consistent with this, week-long SSRI treatment in healthy volunteers has recently been shown to enhance the impact of positive mood inductions on subsequent learning, consistent with the proposal that the delayed mood response relates to how SSRIs influence learning in the world (Michely et al. 2020). Dense sampling is less burdensome for both participants and experimenters when done remotely, and as such, there is great

potential for smartphone-based research to enhance our understanding of drug mechanisms of action in real-life, longitudinal settings. Detailed evaluation of any process that changes over a 4–6-week period due to pharmacological treatment will be hindered by logistical difficulties and potentially prohibitive costs if participants must travel many times to a research site. Smartphone-based data collection greatly reduces the cost of large-scale research on common pharmacological treatments and, combined with experience sampling, allows testing neuroscientific theories about the relationship between drugs taken, real-life events, and symptoms.

As repeated assessment becomes the norm in cognitive studies in mental health, we may also learn that one-shot cognitive testing is less informative than are methods that allow us to estimate variance in performance or average performance over time. This is important because we know that cognitive abilities (as measured through tasks) can vary considerably within an individual, depending on factors such as sleep, stress, and caffeine (Goel et al. 2009, Jarvis 1993, Lieberman et al. 2002), reducing the accuracy of our estimates. Beyond this, variance in performance is of increasing interest to the field. Certain psychiatric populations such as schizophrenia patients exhibit an increased variability in their performance on cognitive tests, in addition to reduced overall performance (Pietrzak et al. 2009). A clinically meaningful observation, experience sampling was used to estimate within-subject variance in mood, and this was found to be an important predictor of future depression status (van de Leemput et al. 2014). The extent to which this applies to cognition is a relatively open question and one which smartphone-based assessments can facilitate.

Digital Phenotyping

The previous section highlighted the potential for regular repeated assessments gathered via smartphone to improve our understanding of how cognitive changes manifest in the real world, change over time, and interact with emotional, social, and physical states. Unfortunately, gathering these data requires high compliance on the part of research participants, which likely introduces systematic bias in sample selection, data completeness, and attrition (Scollon et al. 2003). Other potential issues with this methodology include the facts that reflecting on a behavior or internal state can sometimes alter it and that many cognitive tests carry practice effects. Sensors on smartphones provide additional complementary tools without these limitations in the form of passive data gathering (Harari et al. 2017), which can be used to create so-called digital phenotypes (Insel 2018). Passive data refers to measurements that are gathered automatically without requiring active engagement or submission of data by the research participant. Common measures that can be derived from smartphones include sensor data such as accelerometer; global positioning system (GPS), or light sensors; data pertaining to text messaging, emails, calls, and app use, including social media use; and even microphone or camera data. These data can be used to infer aspects of everyday behavior of interest to researchers in psychiatry, including social engagement, mobility, sleep, and exercise (Cornet & Holden 2018, Mohr et al. 2017).

Sleep disturbance is a core diagnostic feature of depression (APA 2013, Tsuno et al. 2005) that can be difficult to measure retrospectively. Using light sensor and phone use data (Abdullah et al. 2014, Wang et al. 2014), studies have described sleep disturbance in depression, including that sufferers are later to bed, later to rise and more likely to wake at night (Ben-Zeev et al. 2015). Wrist-worn accelerometer data from 91,105 UK Biobank participants showed that depression and bipolar disorder were both associated with disrupted sleep patterns (Lyall et al. 2018). Higher levels of depression are also associated with reductions in GPS-derived metrics of mobility that tap into sedentary aspects of the condition. Individuals with depression visit fewer locations, spend more time at home, and move less through geographic space (Ben-Zeev et al. 2015, Canzian & Musolesi 2015, Saeb et al. 2015), and the opposite is true of individuals drawn from the general

population who have high levels of positive affect and exhibit greater variability in locations visited (Heller et al. 2020). Likewise, happiness has been linked to temporal fluctuations in both exercise and more general physical activity assayed from accelerometer data (Lathia et al. 2017), mirroring findings from a recent large-scale, self-report investigation linking exercise and depression (Chekroud et al. 2018). In terms of social engagement, audio data can be used to quantify conversation frequency and duration, which are both reduced in depression (Wang et al. 2014), and Bluetooth data related to the presence of nearby devices can also act as a proxy for social interactions. Together, these metrics have been utilized to develop prediction/detection tools that might have practical clinical value in the future—for example, in predicting upcoming manic episodes (Abdullah et al. 2016) or relapse in psychosis (Barnett et al. 2018, Ben-Zeev et al. 2017).

In terms of mechanism, however, there exists a major gap between digital phenotyping from passive data and new insights into the changes that occur in the brain that account for these associations. While the cross-sectional association between physical activity and improved mental health is now well established, causality is likely bidirectional (Pinto Pereira et al. 2014), and the neurobiological processes that explain this effect remain poorly understood. Targeting the former issue, smartphone-derived passive assays of activity can allow us to develop directional models in a real-world setting, inferring evidence for causality from the temporal dynamics of events. Crucially, smartphone-based cognitive assessments gathered in tandem have the potential to uncover key brain mediators of these important relationships by providing richer data from tasks designed to probe the specific neural circuits that are believed to be most impacted by mental illness. Future studies should collect passive data, experience sampling, and task performance over time in patients to provide a detailed picture of illness trajectory.

Natural Language/Text Mining

In addition to these indirect forms of digital phenotyping, a major category of passive data that can be gathered in great volume from smartphones concerns language use. The question of what language can tell us about a person's current or future mental state has been of considerable interest in psychiatry for many decades now (Pennebaker et al. 2003). In schizophrenia, speech disturbances like alogia or poverty of speech are well-established diagnostic features (APA 2013), which can to a certain extent be quantified objectively using vocal analysis (Cohen et al. 2014). Recently, a proof of concept has shown this can be done outside of well-controlled settings using videos gathered via smartphone at key points throughout the day (Cohen et al. 2020). These symptoms can be distinguished from other clinical characteristics, such as the flight of ideas seen in mania, which can be accessed through graphical analysis of narratives produced by patients (Mota et al. 2012). Beyond explicit diagnostic features, studies have shown that depressed individuals are more internally focused in their language use, using first-person pronouns like "I" to a greater extent than healthy individuals, both orally and in written word (Bucci & Freedman 1981, Rude et al. 2004). Smartphones provide an excellent source of linguistic data through text messaging and audio data and also from social media posts on third-party apps such as Twitter, Facebook, and Instagram. Language derived from Twitter posts, for example, has been shown to closely mirror that from other more traditional sources in terms of its ability to track depression status (De Choudhury et al. 2013, Reece et al. 2017).

Social media data therefore may present an interesting alternative to traditional ecological momentary assessment (EMA) methods, allowing researchers to assess changes in cognition in tandem with language features and self-reported clinical data over time. Such an approach may prove crucial for developing mechanistic explanations for cross-sectional observations. The advantage here is that rather than asking research participants to regularly complete self-report

questionnaires in an EMA study, microblogging sites hold rich longitudinal archives of not just subtle linguistic features but also semantic content pertaining to users' emotional states, thoughts, and recent events. These data could allow researchers to study longer timescales and in larger samples than can typically be gathered using explicit EMA approaches. For example, archival microblogging data could be used to test whether linguistic features characteristic of a disorder also precipitate the transition into an episode, providing a window into causation and/or early intervention, or in the case of suicide, prevention (Braithwaite et al. 2016). Recently, these data were used to understand how the network dynamics of depression change during episodes of illness. Kelley & Gillan (2020) identified linguistic features characteristic of depression from the tweets of 946 individuals. In a subset of that sample ($N = 286$) who experienced a depressive episode in the 12-month period under study, the authors found that these depression features became more tightly interdependent when a person was ill.

This new methodology for studying changes in clinical features over time is timely and ripe for integration with other neuroscientific tools such as pharmacological interventions and brain imaging. Recent work identifying the potential for more standard forms of EMA to reveal early warning signs for depression, for example, assayed through changes in the autocorrelation, variance, and network connectivity of emotions prior to the onset of episodes (van de Leemput et al. 2014). Although suggestive, findings have been based either on between-subject comparisons (van de Leemput et al. 2014) or on very small samples [$N = 1$ (Wichers et al. 2016)]. Critical transitions into and out of clinical episodes could in theory be examined at a much larger scale (within and between subjects) if self-archived, daily, emotional data from social media are of sufficient quality. Answering some of these questions (e.g., regarding whether or not there is an increase in autocorrelation of depression features prior to an episode) will require particularly dense sampling from social media. This means that for certain questions, only the most frequent posters (posting at least once per day) will contribute data of sufficient granularity. That social media apps are in widespread use on smartphones means that the availability of this sort of data continues to increase and there is considerable scope for custom experimental apps to leverage those time-series data to greatly enhance our understanding of cause and effect with respect to cognitive and neuroscientific markers of mental health and illness.

Enriching Hard Neuroscientific Investigations

A limitation of smartphone-based cognitive neuroscience research is that we cannot simultaneously collect most of the hard measures that are the mainstay of human cognitive neuroscience research. While functional magnetic resonance imaging (fMRI) or positron emission tomography scans are unlikely to ever be collected remotely, and we cannot currently collect saliva or blood samples via smartphone, there are several ways that researchers can bridge this crucial gap. First and foremost, one should, for the most part, view large-scale smartphone studies as a complement to in-person work. The former gathers vast but noisy data, the latter gathers smaller data sets of higher detail and quality. These methods can proceed in tandem (Haworth et al. 2007) but may also occur in series to directly inform one another (Gillan & Seow 2020). For example, Coughlan et al. (2019) utilized the Sea Hero Quest app to develop and test a new measure of spatial navigation in over 27,000 individuals (task depicted in **Figure 3**). Players see a map with several checkpoints (i.e., numbered buoys) to find in a specific order. Players navigate a boat based on numerous sources of information (e.g., memory, environmental cues, tracking their location in space, and total distance travelled) to achieve wayfinding goals, which provides one measure of spatial navigation ability. This large data set allowed them to develop spatial navigation benchmarks that were adjusted for age, education, and gender. Crucially, they then brought this forward to test a

smaller, genetically characterized sample of 60 individuals. They found that their benchmark test was sensitive to a preclinical marker of Alzheimer's disease, apolipoprotein e4 allele (APOE-4), moving from large-scale cognitive phenotyping to a well-defined genetic marker.

Smartphone-based methods have also yielded mechanistic advances of clinical value that were untenable using traditional methods. For example, lab-based pharmacological (Rigoli et al. 2016, Rutledge et al. 2015) and neuroimaging (Chew et al. 2019) studies support the idea that dopamine plays a value-independent role in risk-taking for rewards that can be captured with computational modelling. Natural ageing is associated with a gradual decline in the dopamine system (Bäckman et al. 2006), but effects of ageing on risk-taking are inconsistent (Samanez-Larkin & Knutson 2015), possibly due to the large samples required to identify what are likely to be small effect sizes. Using gamified cognitive testing via smartphone, researchers were able to show that ageing is associated with reduced risk-taking in trials with potential rewards (but not losses) in over 25,000 players of The Great Brain Experiment app (Rutledge et al. 2016). Computational modelling showed that this effect did not depend on the value of the risky option, consistent with lab-based findings with respect to the role of dopamine in risk-taking, suggestive of a potential mechanism for age-related decline. Because the smartphone study's sample size was sufficiently large to detect tiny effect sizes, the lack of association between ageing and decreased risk-taking for potential losses is made even more compelling. In two other games in the app, value-independent reward seeking also decreased with age in a motor decision task requiring participants to make complex motor actions (Chen et al. 2018), and value-independent model parameters predicted information sampling biases in a card game in which participants paid points to flip over cards before making risky decisions (Hunt et al. 2016).

Recently, studies have moved to link data gathered online to fMRI and electroencephalography (EEG). One study used internet-based testing to acquire a large enough sample to define novel self-report transdiagnostic dimensions of impulsivity and compulsivity (Parkes et al. 2019). The weights required to transform responses into individual scores on the impulsivity and compulsivity dimensions were then applied to a smaller sample of diagnosed patients who underwent MRI scanning. The researchers found that these self-reported impulsivity and compulsivity dimensions were associated with different patterns of effective connectivity, while diagnostic information was much less informative. Another study used a similar approach, applying weights from a previously published online study with over 1,400 subjects (Gillan et al. 2016) to characterize the compulsivity levels of just under 200 in-person participants, who underwent EEG while performing a model-based learning task (Seow et al. 2020). This allowed the researchers to probe the underlying neural mechanisms of deficient model-based planning in compulsivity. Using this method, they found evidence that weaker neural representations of state transitions are characteristic of those high in compulsivity, suggesting that previously described deficits might arise from a failure to learn an accurate model of the world. Though illustrative examples, neither study was conducted via smartphone, relying instead on browser-based assessment. This gap was recently filled in a study that found greater diversity in physical location, assessed by geolocation tracking via smartphone, was linked to greater positive affect assessed by experience sampling. Crucially, resting-state fMRI data were collected on roughly half of the subjects ($N = 58$) and revealed that this association was stronger in individuals with greater hippocampal-striatal neural connectivity (Heller et al. 2020). Given the growing number of structural and connectivity neuroimaging studies, which can have more than 1,000 individuals (e.g., Baker et al. 2019), tracking patients with smartphones before and after scanning provides rich data that can be linked to neural measurements.

An even more direct approach than those cited so far concerns the use of mobile EEG devices (Lau-Zhu et al. 2019). Mobile EEG is still a relatively new area of research, but studies have shown that although they have poorer signal-to-noise ratios than traditional systems, reliable, albeit basic,

signals can be gathered from wearable dry-electrode sets (Radüntz 2018). More recently, studies have shown that mobile EEG data can not only be integrated with cognitive tests delivered simultaneously via smartphone (Stopczynski et al. 2014) but also be processed in real time and fed back to the device/user (Blum et al. 2017). This is an exciting prospect, the potential of which was recently exemplified in a smartphone study of learning-related processes and wearable EEG in 10 volunteers (Eldar et al. 2018). Subjects reported their mood four times a day and played a reward learning task twice a day for one week while EEG and heart rate were monitored. The authors found that subjects' neural reward sensitivity, i.e., the extent to which EEG in a session corrected decoded prediction errors, was predictive of later changes in mood measured on the smartphone. This example nicely illustrates the potential for a suite of new investigations via smartphone that can elaborate on candidate mechanisms of future clinical change, both using naturalistic designs and in the context of treatment, relapse monitoring, or even neurofeedback-based interventions. Although they are crucial for building mechanistic models and defining targets for causal manipulation in animal models, it is also important to note that, in many cases, it will not be necessary to incorporate these more direct neuroscientific measures into smartphone-based studies in psychiatry. Smartphone-based approaches can already provide rich information about mental health and illness over time without any direct brain measures. The complexity of psychiatric disorders and the heterogeneity of symptoms suggest that running well-powered neuroscience studies will often lie beyond the resources available to any one clinical neuroscience lab. The best tasks and measures to be utilized by large collaborative neuroscience consortia are those that have already been tested longitudinally in large samples. Increasing the capacity for large-scale, smartphone-based testing should be a key target for researchers, particularly those concerned with clinical translation. Neuroscience-informed, smartphone-based diagnostics and/or interventions will be much less expensive and more scalable than hard measures, which, if successful, will allow for an unprecedent democratization of access to early identification tools, interventions, and more.

Prediction Over Description

Although smartphone-based assessments are well poised to improve our mechanistic descriptions of static states of mental health and illness through rich, multivariate assessment in large samples, one of the most exciting opportunities these methods present is for longitudinal research. With traditional methods, it is simply too difficult to follow enough people for enough time to observe a clinical outcome in sufficient numbers for meaningful analysis. Although largely untapped to date, by virtue of their large sample sizes and the relative ease with which samples can be retained through time, smartphone-based projects are poised to help us to achieve the goal of clinical translation. Sea Hero Quest was an app designed to improve our understanding of spatial navigation on a grand scale and provide new metrics that might provide sensitive markers of Alzheimer's disease in the future. Having achieved over 4.3 million downloads, there is incredible scope for longitudinal and relatively unobtrusive follow-ups to measure cognitive changes 5 or even 10 years after initial sign-up.

On a shorter timescale, a crucial area of research for cognitive neuroscientists seeking clinical translation is treatment prediction—developing methods can determine who is most likely to benefit from an intervention and thereby assist clinical decision-making. It has become increasingly clear that treatment response in psychiatry is highly variable across individuals (Rush et al. 2006). Decades of research investigating potential single-variable markers of treatment response have come up empty handed, and there is growing consensus that success will likely require complex, multivariate modelling approaches such as machine learning (Gillan & Whelan 2017, Rutledge et al. 2019). Machine learning approaches that rely on self-report data exclusively have

shown potential for predicting response to antidepressants in a reanalysis of clinical trial data from over 4,000 patients (Chekroud et al. 2016). Hierarchical clustering of individual symptom items from over 7,000 patients with depression identified three robust symptom clusters that differed in antidepressant response (Chekroud et al. 2017). Excitingly, there are clear indications that such predictions can be enhanced through the addition of cognitive measures (Whelan et al. 2014).

Smartphones offer a new route to convenient longitudinal tracking of symptoms in individuals who have recently started a new treatment. Principal among the opportunities is the growing uptake of internet-based psychological interventions such as internet-based cognitive behavioral therapy (iCBT). While research in antidepressants is to some extent rate limited by the challenge in recruiting individuals who are about to begin treatment, research partnerships with providers of internet-delivered therapies can allow for seamless enrolment of participants in treatment and research that truly scales. Though this has yet to be fully tapped, a recent analysis of an archival iCBT data set with over 50,000 individuals who completed treatment illustrates the potential insightfulness of such partnerships (Chien et al. 2020). Machine learning methods were used to identify subtypes of users, who engaged in different ways with the iCBT tools on offer and ultimately had a more or less successful course of treatment. Coupling these sorts of data with rich cognitive and clinical assessment would help further elucidate which elements of treatment (e.g., self-reflection, supporter interaction, behavioral homework, and psychoeducation) and what treatment durations work best for which individual. In sum, there is significant potential for smartphone-based methodologies to assist in a push toward treatment-focused research that translates complex data sets into individualized predictions of real clinical value (Gillan & Whelan 2017).

Is There Such a Thing as Too Much Data?

Rich and dense data sets combining repeated cognitive testing, experience sampling, self-report clinical questionnaires, and passive data have enormous potential to advance our understanding of mental illness and to make clinically valuable predictions. Collecting these data in individuals for whom neural measurements exist, including neuroimaging-derived structural and functional connectivity, will make possible additional insights into the underlying mechanisms. However, data sets with fewer participants than data points present challenges for robust analysis, and as smartphone-based data sets become larger, this problem can get worse, and external validation becomes increasingly important. Machine learning approaches offer ways to cluster individuals to make clinical predictions, and related dimensionality-reduction methods provide a way of capturing substantial variance in a data set with a smaller number of variables. For example, canonical correlation analysis can identify robust patient clusters, and techniques like L2 regularization can improve performance (Grosenick et al. 2019). Simpler methods such as factor analysis have been used successfully in a number of studies to reduce large sets of self-report questionnaire responses (e.g., 209 items) to three transdiagnostic symptom dimensions that appear to have greater links to underlying biology than extant diagnosis-based summaries of the same data. Compulsivity, one of the dimensions that emerged from this approach, has a specific hallmark where individuals at the higher end of the spectrum reliably have deficits in model-based planning (Gillan et al. 2016). Scores on this compulsivity dimension are dissociable from other dimensions of mental illness characterized by anxiety and depression or social withdrawal, which have their own particular cognitive correlates (Hunter et al. 2019, Rouault et al. 2018, Seow & Gillan 2020).

In complement to these more data-driven methods, theory-driven computational modelling approaches provide a way to exploit knowledge of the generative processes that underlie behavior to efficiently summarize a large amount of data with a small number of parameters (for example, a learning rate and a stochasticity parameter). Parameter estimates can act as an input to data-driven

machine learning approaches, and this combined approach can outperform machine learning approaches alone. For example, classifier performance is higher for simulated agents that differ in learning rates when classifiers are trained on parameters estimated from the data and not directly on the raw data (Huys et al. 2016). The combination of theory-driven and data-driven approaches has great potential to improve understanding and treatment of mental illness, and we see the proliferation of smartphone-based data collection methods as a natural means of facilitating this. Care must be taken to avoid algorithmic biases that have been shown to inherit biases present in training data sets. For example, algorithms trained using internet-based language corpora inherit common gender and race biases (Caliskan et al. 2017). Furthermore, validation with independent samples is essential for reaching conclusions that are likely to generalize (Rutledge et al. 2019).

CONCLUSION

Smartphones have great potential to increase the volume of data available to researchers in psychiatry by multiple orders of magnitude. This is important because it is becoming increasingly clear that univariate effect sizes in cognitive neuropsychiatry are likely small and, in order to be predictive, will need contextualization using concurrent measurement of between-subject variables like age and education or within-subject state variables like recent sleep quality or current stress. Computational tools are emerging that can help us make sense of these vast data sets and to link findings to research on the underlying neural mechanisms. At this point, standard approaches are unlikely to produce the new treatments needed for substantial improvements in our understanding and treatment of mental illnesses. Smartphones complement existing neuroscientific approaches and may be an important tool for getting the data we need as we grapple with the complexity of mental well-being.

DISCLOSURE STATEMENT

The authors are not aware of any affiliations, memberships, funding, or financial holdings that might be perceived as affecting the objectivity of this review.

ACKNOWLEDGMENTS

LITERATURE CITED

Abdullah S, Matthews M, Frank E, Doherty G, Gay G, Choudhury T. 2016. Automatic detection of social rhythms in bipolar disorder. *J. Am. Med. Inform. Assoc.* 23(3):538–43

Abdullah S, Matthews M, Murnane EL, Gay G, Choudhury T. 2014. *Towards circadian computing: "Early to bed and early to rise" makes some of us unhealthy and sleep deprived*. Paper presented at UBICOMP, Seattle, WA

APA (Am. Psychiatr. Assoc.). 2013. *Diagnostic and Statistical Manual of Mental Disorders: DSM-5*. Washington, DC: Am. Psychiatr. Publ. 5th ed.

Bäckman L, Nyberg L, Lindenberger U, Li SC, Farde L. 2006. The correlative triad among aging, dopamine, and cognition: current status and future prospects. *Neurosci. Biobehav. Rev.* 30(6):791–807

Baker JT, Dillon DG, Patrick LM, Roffman JL, Brady RO, et al. 2019. Functional connectomics of affective and psychotic pathology. *PNAS* 116(18):9050–59

Barnett I, Torous J, Staples P, Sandoval L, Keshavan M, Onnela JP. 2018. Relapse prediction in schizophrenia through digital phenotyping: a pilot study. *Neuropsychopharmacology* 43(8):1660–66

Bedder R, Vaghi MM, Dolan RJ, Rutledge R. 2020. Risk taking for potential losses but not gains increases with time of day. PsyArXiv. **https://doi.org/10.31234/osf.io/3qdnx**

Ben-Zeev D, Brian R, Wang R, Wang W, Campbell AT, et al. 2017. CrossCheck: integrating self-report, behavioral sensing, and smartphone use to identify digital indicators of psychotic relapse. *Psychiatr. Rehabil. J.* 40(3):266–75

Ben-Zeev D, Scherer EA, Wang R, Xie H, Campbell AT. 2015. Next-generation psychiatric assessment: using smartphone sensors to monitor behavior and mental health. *Psychiatr. Rehabil. J.* 38(3):218–26

Bickel WK, Jarmolowicz DP, Mueller ET, Koffarnus MN, Gatchalian KM. 2012. Excessive discounting of delayed reinforcers as a trans-disease process contributing to addiction and other disease-related vulnerabilities: emerging evidence. *Pharmacol. Ther.* 134(3):287–97

Blain B, Rutledge RB. 2020. Momentary subjective well-being depends on learning and not reward. *eLife* 9:e57977

Blum S, Debener S, Emkes R, Volkening N, Fudickar S, Bleichner MG. 2017. EEG recording and online signal processing on Android: a multiapp framework for brain-computer interfaces on smartphone. *Biomed. Res. Int.* 2017:3072870

Braithwaite SR, Giraud-Carrier C, West J, Barnes MD, Hanson CL. 2016. Validating machine learning algorithms for Twitter data against established measures of suicidality. *JMIR Ment. Health* 3(2):e21

Brown HR, Zeidman P, Smittenaar P, Adams RA, McNab F, et al. 2014. Crowdsourcing for cognitive science—the utility of smartphones. *PLOS ONE* 9(7):e100662

Browning M, Carter C, Chatham C, Den Ouden H, Gillan CM, et al. 2020. Realising the clinical potential of computational psychiatry: report from the Banbury Centre meeting, February 2019. *Biol. Psychiatry* 88(2):e5–10

Bucci W, Freedman N. 1981. The language of depression. *Bull. Menninger Clin.* 45(4):334–58

Buhrmester M, Kwang T, Gosling SD. 2011. Amazon's Mechanical Turk: a new source of inexpensive, yet high-quality, data? *Perspect. Psychol. Sci.* 6(1):3–5

Button KS, Ioannidis JP, Mokrysz C, Nosek BA, Flint J, et al. 2013. Power failure: why small sample size undermines the reliability of neuroscience. *Nat. Rev. Neurosci.* 14(5):365–76

Caliskan A, Bryson JJ, Narayanan A. 2017. Semantics derived automatically from language corpora contain human-like biases. *Science* 356(6334):183–86

Canzian L, Musolesi M. 2015. *Trajectories of depression: unobtrusive monitoring of depressive states by means of smartphone mobility traces analysis*. Paper presented at UBICOMP, Osaka, Japan

Casler K, Bickel L, Hackett E. 2013. Separate but equal? A comparison of participants and data gathered via Amazon's MTurk, social media, and face-to-face behavioral testing. *Comput. Hum. Behav.* 29(6):2156–60

Chandler J, Paolacci G. 2017. Lie for a dime: when most prescreening responses are honest but most study participants are impostors. *Soc. Psychol. Personal. Sci.* 8(5):500–8

Chandler J, Shapiro D. 2016. Conducting clinical research using crowdsourced convenience samples. *Annu. Rev. Clin. Psychol.* 12:53–81

Chekroud AM, Gueorguieva R, Krumholz HM, Trivedi MH, Krystal JH, McCarthy G. 2017. Reevaluating the efficacy and predictability of antidepressant treatments: a symptom clustering approach. *JAMA Psychiatry* 74(4):370–78

Chekroud AM, Zotti RJ, Shehzad Z, Gueorguieva R, Johnson MK, et al. 2016. Cross-trial prediction of treatment outcome in depression: a machine learning approach. *Lancet Psychiatry* 3(3):243–50

Chekroud SR, Gueorguieva R, Zheutlin AB, Paulus M, Krumholz HM, et al. 2018. Association between physical exercise and mental health in 1.2 million individuals in the USA between 2011 and 2015: a cross-sectional study. *Lancet Psychiatry* 5(9):739–46

Chen X, Rutledge RB, Brown HR, Dolan RJ, Bestmann S, Galea JM. 2018. Age-dependent Pavlovian biases influence motor decision-making. *PLOS Comput. Biol.* 14(7):e1006304

Chew B, Hauser TU, Papoutsi M, Magerkurth J, Dolan RJ, Rutledge RB. 2019. Endogenous fluctuations in the dopaminergic midbrain drive behavioral choice variability. *PNAS* 116(37):18732–37

Chien I, Enrique A, Palacios J, Regan T, Keegan D, et al. 2020. A machine learning approach to understanding patterns of engagement with internet-delivered mental health interventions. *JAMA Netw. Open* 3(7):e2010791

Clarke DE, Narrow WE, Regier DA, Kuramoto SJ, Kupfer DJ, et al. 2013. DSM-5 field trials in the United States and Canada, Part I: study design, sampling strategy, implementation, and analytic approaches. *Am. J. Psychiatry* 170(1):43–58

Cohen AS, Cowan T, Le TP, Schwartz EK, Kirkpatrick B, et al. 2020. Ambulatory digital phenotyping of blunted affect and alogia using objective facial and vocal analysis: proof of concept. *Schizophr. Res.* 220:141–46

Cohen AS, Mitchell KR, Elvevåg B. 2014. What do we really know about blunted vocal affect and alogia? A meta-analysis of objective assessments. *Schizophr. Res.* 159(2–3):533–38

Cornet VP, Holden RJ. 2018. Systematic review of smartphone-based passive sensing for health and wellbeing. *J. Biomed. Inform.* 77:120–32

Coughlan G, Coutrot A, Khondoker M, Minihane AM, Spiers H, Hornberger M. 2019. Toward personalized cognitive diagnostics of at-genetic-risk Alzheimer's disease. *PNAS* 116(19):9285–92

Coutrot A, Silva R, Manley E, de Cothi W, Sami S, et al. 2018. Global determinants of navigation ability. *Curr. Biol.* 28(17):2861–66.e4

Crump MJ, McDonnell JV, Gureckis TM. 2013. Evaluating Amazon's Mechanical Turk as a tool for experimental behavioral research. *PLOS ONE* 8(3):e57410

Daniel-Watanabe L, McLaughlin M, Gormley S, Robinson OJ. 2020. Association between a directly translated cognitive measure of negative bias and self-reported psychiatric symptoms. *Biol. Psychiatry: Cogn. Neurosci. Neuroimaging*. In press

Davidson LL, Heinrichs RW. 2003. Quantification of frontal and temporal lobe brain-imaging findings in schizophrenia: a meta-analysis. *Psychiatry Res.* 122(2):69–87

Daw ND, Gershman SJ, Seymour B, Dayan P, Dolan RJ. 2011. Model-based influences on humans' choices and striatal prediction errors. *Neuron* 69(6):1204–15

Dayan P, Huys QJ. 2008. Serotonin, inhibition, and negative mood. *PLOS Comput. Biol.* 4(2):e4

De Choudhury M, Gamon M, Counts S, Horvitz E. 2013. *Predicting depression via social media*. Paper presented at the 7th International AAAI Conference on Weblogs and Social Media, Cambridge, MA

Eldar E, Roth C, Dayan P, Dolan RJ. 2018. Decodability of reward learning signals predicts mood fluctuations. *Curr. Biol.* 28(9):1433–39.e7

Farrell MS, Werge T, Sklar P, Owen MJ, Ophoff RA, et al. 2015. Evaluating historical candidate genes for schizophrenia. *Mol. Psychiatry* 20(5):555–62

Flint J, Kendler KS. 2014. The genetics of major depression. *Neuron* 81(3):484–503

Fluharty M, Taylor AE, Grabski M, Munafò MR. 2017. The association of cigarette smoking with depression and anxiety: a systematic review. *Nicotine Tob. Res.* 19(1):3–13

Ford DE, Kamerow DB. 1989. Epidemiologic study of sleep disturbances and psychiatric disorders. An opportunity for prevention? *JAMA* 262(11):1479–84

Fried EI, Nesse RM. 2015. Depression is not a consistent syndrome: an investigation of unique symptom patterns in the STAR*D study. *J. Affect. Disord.* 172:96–102

Fuller PM, Gooley JJ, Saper CB. 2006. Neurobiology of the sleep-wake cycle: sleep architecture, circadian regulation, and regulatory feedback. *J. Biol. Rhythms* 21:482–93

Germine L, Nakayama K, Duchaine BC, Chabris CF, Chatterjee G, Wilmer JB. 2012. Is the Web as good as the lab? Comparable performance from Web and lab in cognitive/perceptual experiments. *Psychon. Bull. Rev.* 19(5):847–57

Gillan CM, Daw ND. 2016. Taking psychiatry research online. *Neuron* 91(1):19–23

Gillan CM, Fineberg NA, Robbins TW. 2017. A trans-diagnostic perspective on obsessive-compulsive disorder. *Psychol. Med.* 47(9):1528–48

Gillan CM, Kalanthroff E, Evans M, Weingarden HM, Jacoby RJ, et al. 2019. Comparison of the association between goal-directed planning and self-reported compulsivity versus obsessive-compulsive disorder diagnosis. *JAMA Psychiatry* 77(1):1–10

Gillan CM, Kosinski M, Whelan R, Phelps EA, Daw ND. 2016. Characterizing a psychiatric symptom dimension related to deficits in goal-directed control. *eLife* 5:e11305

Gillan CM, Seow TXF. 2020. Carving out new transdiagnostic dimensions for research in mental health. *Biol. Psychiatry: Cogn. Neurosci. Neuroimaging* 5(10):932–34

Gillan CM, Whelan R. 2017. What big data can do for treatment in psychiatry. *Curr. Opin. Behav. Sci.* 18:34–42

Goel N, Rao H, Durmer JS, Dinges DF. 2009. Neurocognitive consequences of sleep deprivation. *Semin. Neurol.* 29(4):320–39

Goldman N. 1994. Social factors and health: the causation-selection issue revisited. *PNAS* 91(4):1251–55

Goodman JK, Cryder CE, Cheema A. 2013. Data collection in a flat world: the strengths and weaknesses of Mechanical Turk samples. *J. Behav. Decis. Making* 26:213–24

Grosenick L, Shi TC, Gunning FM, Dubin MJ, Downar J, Liston C. 2019. Functional and optogenetic approaches to discovering stable subtype-specific circuit mechanisms in depression. *Biol. Psychiatry Cogn. Neurosci. Neuroimaging* 4(6):554–66

Harari GM, Müller SR, Aung MS, Rentfrow P. 2017. Smartphone sensing methods for studying behavior in everyday life. *Curr. Opin. Behav. Sci.* 18:83–90

Haslam N, Holland E, Kuppens P. 2012. Categories versus dimensions in personality and psychopathology: a quantitative review of taxometric research. *Psychol. Med.* 42(5):903–20

Haworth CM, Harlaar N, Kovas Y, Davis OS, Oliver BR, et al. 2007. Internet cognitive testing of large samples needed in genetic research. *Twin Res. Hum. Genet.* 10(4):554–63

Hedge C, Powell G, Sumner P. 2018. The reliability paradox: why robust cognitive tasks do not produce reliable individual differences. *Behav. Res. Methods* 50(3):1166–86

Heller AS, Shi TC, Ezie CEC, Reneau TR, Baez LM, et al. 2020. Association between real-world experiential diversity and positive affect relates to hippocampal-striatal functional connectivity. *Nat. Neurosci.* 23(7):800–4

Henrich J, Heine SJ, Norenzayan A. 2010. The weirdest people in the world? *Behav. Brain Sci.* 33(2–3):61–83

Hoogman M, Muetzel R, Guimaraes JP, Shumskaya E, Mennes M, et al. 2019. Brain imaging of the cortex in ADHD: a coordinated analysis of large-scale clinical and population-based samples. *Am. J. Psychiatry* 176(7):531–42

Hunt LT, Rutledge RB, Malalasekera WM, Kennerley SW, Dolan RJ. 2016. Approach-induced biases in human information sampling. *PLOS Biol.* 14(11):e2000638

Hunter LE, Meer EA, Gillan CM, Hsu M, Daw ND. 2019. Excessive deliberation in social anxiety. bioRxiv 522433. **https://doi.org/10.1101/522433**

Huys QJ, Maia TV, Frank MJ. 2016. Computational psychiatry as a bridge from neuroscience to clinical applications. *Nat. Neurosci.* 19(3):404–13

Insel TR. 2018. Digital phenotyping: a global tool for psychiatry. *World Psychiatry* 17(3):276–77

Jané-Llopis E, Matytsina I. 2006. Mental health and alcohol, drugs and tobacco: a review of the comorbidity between mental disorders and the use of alcohol, tobacco and illicit drugs. *Drug Alcohol Rev.* 25(6):515–36

Jarvis MJ. 1993. Does caffeine intake enhance absolute levels of cognitive performance? *Psychopharmacology* 110(1–2):45–52

Kalaria RN, Maestre GE, Arizaga R, Friedland RP, Galasko D, et al. 2008. Alzheimer's disease and vascular dementia in developing countries: prevalence, management, and risk factors. *Lancet Neurol.* 7(9):812–26

Kapur S, Phillips AG, Insel TR. 2012. Why has it taken so long for biological psychiatry to develop clinical tests and what to do about it? *Mol. Psychiatry* 17(12):1174–79

Kelley S, Gillan C. 2020. Within-subject changes in network connectivity occur during an episode of depression: evidence from a longitudinal analysis of social media posts. PsyArXiv. **https://doi.org/10.31234/osf.io/6h52d**

Kelly JR, Kennedy PJ, Cryan JF, Dinan TG, Clarke G, Hyland NP. 2015. Breaking down the barriers: the gut microbiome, intestinal permeability and stress-related psychiatric disorders. *Front. Cell Neurosci.* 9:392

Kessler RC, Davis CG, Kendler KS. 1997. Childhood adversity and adult psychiatric disorder in the US National Comorbidity Survey. *Psychol. Med.* 27(5):1101–19

Killingsworth MA, Gilbert DT. 2010. A wandering mind is an unhappy mind. *Science* 330(6006):932

Lathia N, Sandstrom GM, Mascolo C, Rentfrow PJ. 2017. Happier people live more active lives: using smartphones to link happiness and physical activity. *PLOS ONE* 12(1):e0160589

Lau-Zhu A, Lau MPH, McLoughlin G. 2019. Mobile EEG in research on neurodevelopmental disorders: opportunities and challenges. *Dev. Cogn. Neurosci.* 36:100635

Lieberman HR, Tharion WJ, Shukitt-Hale B, Speckman KL, Tulley R. 2002. Effects of caffeine, sleep loss, and stress on cognitive performance and mood during U.S. Navy SEAL training. Sea-Air-Land. *Psychopharmacology* 164(3):250–61

Linn MC, Petersen AC. 1985. Emergence and characterization of sex differences in spatial ability: a meta-analysis. *Child Dev*. 56(6):1479–98

Lipszyc J, Schachar R. 2010. Inhibitory control and psychopathology: a meta-analysis of studies using the stop signal task. *J. Int. Neuropsychol. Soc.* 16(6):1064–76

Lorant V, Deliège D, Eaton W, Robert A, Philippot P, Ansseau M. 2003. Socioeconomic inequalities in depression: a meta-analysis. *Am. J. Epidemiol.* 157(2):98–112

Lyall LM, Wyse CA, Graham N, Ferguson A, Lyall DM, et al. 2018. Association of disrupted circadian rhythmicity with mood disorders, subjective wellbeing, and cognitive function: a cross-sectional study of 91 105 participants from the UK Biobank. *Lancet Psychiatry* 5(6):507–14

MacKerron G, Mourato A. 2013. Happiness is greater in natural environments. *Glob. Environ. Change* 23(5):992–1000

Marek S, Tervo-Clemmens B, Calabro FJ, Montez DF, Kay BP, et al. 2020. Towards reproducible brain-wide association studies. bioRxiv 2020.08.21.257758. **https://doi.org/10.1101/2020.08.21.257758**

McNab F, Zeidman P, Rutledge RB, Smittenaar P, Brown HR, et al. 2015. Age-related changes in working memory and the ability to ignore distraction. *PNAS* 112(20):6515–18

Michely J, Eldar E, Martin IM, Dolan RJ. 2020. A mechanistic account of serotonin's impact on mood. *Nat. Commun.* 11(1):2335

Mohr DC, Zhang M, Schueller SM. 2017. Personal sensing: understanding mental health using ubiquitous sensors and machine learning. *Annu. Rev. Clin. Psychol.* 13:23–47

Mota NB, Vasconcelos NA, Lemos N, Pieretti AC, Kinouchi O, et al. 2012. Speech graphs provide a quantitative measure of thought disorder in psychosis. *PLOS ONE* 7(4):e34928

Müller VI, Cieslik EC, Serbanescu I, Laird AR, Fox PT, Eickhoff SB. 2017. Altered brain activity in unipolar depression revisited: meta-analyses of neuroimaging studies. *JAMA Psychiatry* 74(1):47–55

Nussenbaum K, Scheuplein M, Phaneuf CV, Evans MD, Hartley CA. 2020. Moving developmental research online: comparing in-lab and web-based studies of model-based reinforcement learning. OSF Preprints. **https://doi.org/10.1525/collabra.17213**

O'Neil A, Quirk SE, Housden S, Brennan SL, Williams LJ, et al. 2014. Relationship between diet and mental health in children and adolescents: a systematic review. *Am. J. Public Health* 104(10):e31–42

Orban C, Kong R, Li J, Chee MW, Yeo BT. 2020. Time of day is associated with paradoxical reductions in global signal fluctuation and functional connectivity. *PLOS Biol*. 18(2):e3000602

Parkes L, Tiego J, Aquino K, Braganza L, Chamberlain SR, et al. 2019. Transdiagnostic variations in impulsivity and compulsivity in obsessive-compulsive disorder and gambling disorder correlate with effective connectivity in cortical-striatal-thalamic-cortical circuits. *Neuroimage* 202:116070

Paykel ES, Abbott R, Jenkins R, Brugha TS, Meltzer H. 2000. Urban-rural mental health differences in great Britain: findings from the national morbidity survey. *Psychol. Med.* 30(2):269–80

Pennebaker JW, Mehl MR, Niederhoffer KG. 2003. Psychological aspects of natural language use: our words, our selves. *Annu. Rev. Psychol.* 54:547–77

Pietrzak RH, Snyder PJ, Jackson CE, Olver J, Norman T, et al. 2009. Stability of cognitive impairment in chronic schizophrenia over brief and intermediate re-test intervals. *Hum. Psychopharmacol.* 24(2):113–21

Pinto Pereira SM, Geoffroy MC, Power C. 2014. Depressive symptoms and physical activity during 3 decades in adult life: bidirectional associations in a prospective cohort study. *JAMA Psychiatry* 71(12):1373–80

Pooseh S, Bernhardt N, Guevara A, Huys QJM, Smolka MN. 2018. Value-based decision-making battery: a Bayesian adaptive approach to assess impulsive and risky behavior. *Behav. Res. Methods* 50(1):236–49

Pringle A, Browning M, Cowen PJ, Harmer CJ. 2011. A cognitive neuropsychological model of antidepressant drug action. *Prog. Neuropsychopharmacol. Biol. Psychiatry* 35(7):1586–92

Radüntz T. 2018. Signal quality evaluation of emerging EEG devices. *Front. Physiol.* 9:98

Reece AG, Reagan AJ, Lix KLM, Dodds PS, Danforth CM, Langer EJ. 2017. Forecasting the onset and course of mental illness with Twitter data. *Sci. Rep.* 7(1):13006

Regier DA, Narrow WE, Clarke DE, Kraemer HC, Kuramoto SJ, et al. 2013. DSM-5 field trials in the United States and Canada, Part II: test-retest reliability of selected categorical diagnoses. *Am. J. Psychiatry* 170(1):59–70

Richard A, Rohrmann S, Vandeleur CL, Schmid M, Barth J, Eichholzer M. 2017. Loneliness is adversely associated with physical and mental health and lifestyle factors: results from a Swiss national survey. *PLOS ONE* 12(7):e0181442

Rigoli F, Rutledge RB, Dayan P, Dolan RJ. 2016. The influence of contextual reward statistics on risk preference. *Neuroimage* 128:74–84

Rouault M, Seow T, Gillan CM, Fleming SM. 2018. Psychiatric symptom dimensions are associated with dissociable shifts in metacognition but not task performance. *Biol. Psychiatry* 84(6):443–51

Rude S, Gortner EM, Pennebaker J. 2004. Language use of depressed and depression-vulnerable college students. *Cogn. Emot.* 18(8):1121–33

Rush AJ, Trivedi MH, Wisniewski SR, Nierenberg AA, Stewart JW, et al. 2006. Acute and longer-term outcomes in depressed outpatients requiring one or several treatment steps: a STAR*D report. *Am. J. Psychiatry* 163(11):1905–17

Rutledge RB, Chekroud AM, Huys QJ. 2019. Machine learning and big data in psychiatry: toward clinical applications. *Curr. Opin. Neurobiol.* 55:152–59

Rutledge RB, Moutoussis M, Smittenaar P, Zeidman P, Taylor T, et al. 2017. Association of neural and emotional impacts of reward prediction errors with major depression. *JAMA Psychiatry* 74(8):790–97

Rutledge RB, Skandali N, Dayan P, Dolan RJ. 2014. A computational and neural model of momentary subjective well-being. *PNAS* 111(33):12252–57

Rutledge RB, Skandali N, Dayan P, Dolan RJ. 2015. Dopaminergic modulation of decision making and subjective well-being. *J. Neurosci.* 35(27):9811–22

Rutledge RB, Smittenaar P, Zeidman P, Brown HR, Adams RA, et al. 2016. Risk taking for potential reward decreases across the lifespan. *Curr. Biol.* 26(12):1634–39

Saeb S, Zhang M, Karr CJ, Schueller SM, Corden ME, et al. 2015. Mobile phone sensor correlates of depressive symptom severity in daily-life behavior: an exploratory study. *J. Med. Internet Res.* 17(7):e175

Samanez-Larkin GR, Knutson B. 2015. Decision making in the ageing brain: changes in affective and motivational circuits. *Nat. Rev. Neurosci.* 16(5):278–89

Scollon CN, Kim-Prieto C, Diener E. 2003. Experience sampling: promises and pitfalls, strengths and weaknesses. *J. Happiness Stud.* 4(1):5–34

Seow TXF, O'Connell R, Gillan CM. 2020. Model-based learning deficits in compulsivity are linked to faulty representations of task structure. bioRxiv 2020.06.11.147447. **https://doi.org/10.1101/2020.06.11.147447**

Seow TXF, Gillan CM. 2020. Transdiagnostic phenotyping reveals a host of metacognitive deficits implicated in compulsivity. *Sci. Rep.* 10:2883

Shapiro DN, Chandler J, Mueller PA. 2013. Using Mechanical Turk to study clinical populations. *Clin. Psychol. Sci.* 1(2):213–20

Smallwood J, Fitzgerald A, Miles LK, Phillips LH. 2009. Shifting moods, wandering minds: Negative moods lead the mind to wander. *Emotion* 9(2):271–76

Stewart N, Chandler J, Paolacci G. 2017. Crowdsourcing samples in cognitive science. *Trends Cogn. Sci.* 21(10):736–48

Stopczynski A, Stahlhut C, Larsen JE, Petersen MK, Hansen LK. 2014. The smartphone brain scanner: a portable real-time neuroimaging system. *PLOS ONE* 9(2):e86733

Taquet M, Quoidbach J, de Montjoye Y-A, Desseilles M, Gross JJ. 2016. Hedonism and the choice of everyday activities. *PNAS* 113:9769–73

Taquet M, Quoidbach J, Gross JJ, Saunders KEA, Goodwin GM. 2020. Mood homeostasis, low mood, and history of depression in 2 large population samples. *JAMA Psychiatry* 77(9):944–51

Triantafillou S, Saeb S, Lattie EG, Mohr DC, Kording KP. 2019. Relationship between sleep quality and mood: ecological momentary assessment study. *JMIR Ment. Health* 6(3):e12613

Tsuno N, Besset A, Ritchie K. 2005. Sleep and depression. *J. Clin. Psychiatry* 66(10):1254–69

van de Leemput IA, Wichers M, Cramer AO, Borsboom D, Tuerlinckx F, et al. 2014. Critical slowing down as early warning for the onset and termination of depression. *PNAS* 111(1):87–92

Villano WJ, Otto AR, Ezie CEC, Gillis R, Heller AS. 2020. Temporal dynamics of real-world emotion are more strongly linked to prediction error than outcome. *J. Exp. Psychol. Gen.* 149(9):1755–66

Voon V, Derbyshire K, Rück C, Irvine MA, Worbe Y, et al. 2014. Disorders of compulsivity: a common bias towards learning habits. *Mol. Psychiatry* 20:345–52

Wang R, Chen F, Chen Z, Li T, Harari G, et al. 2014. StudentLife: assessing mental health, academic performance and behavioral trends of college students using smartphones. In *Proceedings of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing*, pp. 3–14. New York: Assoc. Comput. Mach.

Weitzman ER. 2004. Poor mental health, depression, and associations with alcohol consumption, harm, and abuse in a national sample of young adults in college. *J. Nerv. Ment. Dis.* 192(4):269–77

Whelan R, Watts R, Orr CA, Althoff RR, Artiges E, et al. 2014. Neuropsychosocial profiles of current and future adolescent alcohol misusers. *Nature* 512(7513):185–89

Wichers M, Groot PC, Psychosystems ESM Group, Group EWS. 2016. Critical slowing down as a personalized early warning signal for depression. *Psychother. Psychosom.* 85(2):114–16

Widge AS, Bilge MT, Montana R, Chang W, Rodriguez CI, et al. 2019. Electroencephalographic biomarkers for treatment response prediction in major depressive illness: a meta-analysis. *Am. J. Psychiatry* 176(1):44–56

Zung WW. 1965. A self-rating depression scale. *Arch. Gen. Psychiatry* 12:63–70