ANNUAL REVIEWS

Annual Review of Clinical Psychology

Machine Learning and the Digital Measurement of Psychological Health

Isaac R. Galatzer-Levy[1,2] and Jukka-Pekka Onnela[3]

[1]Department of Psychiatry, New York University Grossman School of Medicine, New York, NY, USA; email: isaac.galatzer-levy@nyumc.org

[2]Current affiliation: Google LLC, Mountain View, California, USA

[3]Department of Biostatistics, Harvard T.H. Chan School of Public Health, Boston, Massachusetts, USA

Annu. Rev. Clin. Psychol. 2023. 19:133–54

The *Annual Review of Clinical Psychology* is online at clinpsy.annualreviews.org

https://doi.org/10.1146/annurev-clinpsy-080921-073212

Keywords

digital phenotyping, machine learning, digital biomarkers, psychometrics

Abstract

Since its inception, the discipline of psychology has utilized empirical epistemology and mathematical methodologies to infer psychological functioning from direct observation. As new challenges and technological opportunities emerge, scientists are once again challenged to define measurement paradigms for psychological health and illness that solve novel problems and capitalize on new technological opportunities. In this review, we discuss the theoretical foundations of and scientific advances in remote sensor technology and machine learning models as they are applied to quantify psychological functioning, draw clinical inferences, and chart new directions in treatment.

Contents

INTRODUCTION

What is the difference between coping with trauma and developing posttraumatic stress disorder (PTSD) after a horrific experience? What does it mean to say that a medication treats depression? Embedded in these questions are deeper questions about the true nature and constructed definitions of psychological health and illness. Differentiating psychological health from illness represents a unique challenge that researchers have struggled to address for more than a century. Indeed, symptoms of psychiatric disorders—such as low mood, arousal, psychomotor retardation, sleep disturbances, inattention, and intrusive thoughts—are common. They do not differentiate health from pathology and do not individually differentiate between psychiatric disorders (Kessler et al. 2005).

Over the past century, there have been numerous paradigm shifts and formal revisions to the definition of psychological health and illness. These revisions have emerged to address new challenges and opportunities for research and treatment that could not be addressed using prior scientific methodologies or technologies (Casey et al. 2013, Galatzer-Levy & Galatzer-Levy 2007, Insel et al. 2010, Spitzer et al. 1978). Across generations, researchers have borrowed from measurement theory, experimental methods, computer science, and data science to improve measurements and fit them to address emerging scientific questions and societal problems.

In this review, we chart the evolution in the measurement of psychological functioning in the context of emerging approaches to assess behavioral and physiological states from remotely embedded sensors and machine learning algorithms. Ultimately, we aim to provide a framework to understand how digital measurements can bridge basic and applied clinical science by providing the precision and objectivity of direct and continuous assessments of key outputs of the brain (behavior and physiology; LeDoux 2012) using technologies that are embedded in people's real-world activities. The ultimate goal is to directly measure and understand psychological health as it unfolds across time and real-world contexts (Hitchcock et al. 2022).

Machine learning algorithms: a class of algorithms that aim to learn one or more functions to represent the relationships between variables in one set of data that are reproducible in new instances

FOUNDATIONS OF CLINICAL MEASUREMENT

Since its inception, psychology has utilized direct and measurable observation to infer underlying psychological functions from behavior and physiology (Ayer 1959, Galatzer-Levy & Galatzer-Levy 2007, James 1890, Wundt & Judd 1902). Clinical researchers extended this framework with the aim of identifying behavioral and physiological parameters that were rare and debilitating.

While common patterns of abnormal behavior and physiology were identified soon after the emergence of this field of inquiry, the methodology by which these indicators are defined and measured has evolved predictably in line with technological innovation. Each innovation has added more specificity to measurement, has provided deeper and more specific knowledge, and has facilitated new approaches to diagnosis and treatment.

History and Aims of Clinical Measurement

The foundations of clinical measurement as an empirical science emerged primarily from the work of Emil Kraepelin and his collaborators. Kraepelin is credited with aligning psychiatry with the methods of psychological science and is widely regarded as the father of scientific psychiatry and pharmacology. Observations made in patients being treated in asylums demonstrated to Kraepelin that there is no unique class of behaviors and actions in people with even the most severe forms of mental illness. Instead, these characteristics (e.g., flattened affect, agitation, lethargy, alogia, catatonia, hallucinations) can occur to some degree in anyone. Kraepelin observed that they are meaningful only when they cluster at a cross section and over time because they can be provoked by many potential precipitants, such as alcohol, poisoning, sleep deprivation, or not-yet-understood internal mechanisms (Kraepelin 1915). This conceptual framework rapidly gained prominence when lithium, a medication that was developed to treat gout, was introduced in 1970 as a treatment for what was then termed bipolar mania (Shorter 2009). The dramatic improvement in functioning of people with bipolar disorder indicated that mental disorders may be similar to other chronic health conditions like high blood pressure, which can be managed by medication. In this context, clear symptom and treatment response definitions became essential to clinical research and care (Galatzer-Levy & Galatzer-Levy 2007). This research provided a framework to conceptualize psychological health and abnormality in terms of longitudinal clusters across time and context that were targets for both measurement and remediation.

Kraepelin's diagnostic scheme provided a framework but lacked definitions that were specific enough to provide reproducible methods for diagnosis and treatment more broadly. To fill this void, Spitzer and colleagues introduced a new framework for assessment and diagnosis, termed the Research Diagnostic Criteria (RDC) (Spitzer et al. 1978). In this framework, ultimately codified in the *Diagnostic and Statistical Manual of Mental Disorders* (DSM) (Spitzer et al. 1975), patients were clinically evaluated systematically, through the use of checklist items that could be algorithmically sorted to determine the correct diagnosis (Endicott et al. 1975, Spitzer & Endicott 1968). This approach had a rapid and widespread influence on definitions of mental health illnesses through the proliferation of a simple and reliable measurement paradigm. The RDC revolution allowed screening, diagnosis, and treatment to be delivered in a standardized manner by creating consistent, reliable, and reproducible methods that could be generalized across research into populations (Kessler et al. 2005, Manea et al. 2015), mechanisms, and treatment (Boesen et al. 2021, Hyman 2010).

The Emergence of New Measurement Paradigms

High comorbidity between disorders (Kessler 1994, Kessler et al. 2005), poor mapping of diagnosis to treatment (Carter et al. 2012), and failure to separate diagnoses on the basis of genetic or brain-based mechanisms (Goodkind et al. 2015, O'Donovan 2015) refuted Spitzer and colleagues' hypothesis that validity of clinical definitions and underlying mechanisms would result from first establishing reliability in measurement (Spitzer et al. 1978). These empirical limitations were ultimately attributed to the complex rules around classification, which produced heterogeneity within diagnoses. Diagnoses were highly heterogeneous because they treated all symptoms as interchangeable and of equal weight, in order to create simple rules for high-dimensional cutoffs

that could be scored by hand. This approach ultimately obfuscated individual symptoms and their severity. For example, there are 3,024 combinations of symptom presentation across the range of symptoms and symptom severity in the Patient Health Questionnaire–9 measure of depression, as well as 636,120 combinations of the PTSD diagnosis (Galatzer-Levy & Bryant 2013).

This hidden heterogeneity makes this measurement paradigm untenable in the context of increasingly focused mechanistic research. For example, DSM criteria failed to differentiate between traumatic and healthy responses to trauma (Harvey & Bryant 1998, Marshall et al. 1999). As another example, the timescale of measurement, such as the requirement that symptoms persist for 2 weeks in the context of major depressive disorder, are not suitable for measuring the effects of new classes of rapid-acting antidepressants (e.g., ketamine, psilocybin) that purport to cause shifts in symptoms in minutes rather than weeks (Cusin et al. 2010). These limitations made RDC a blunt measurement tool for increasingly sophisticated measurement requirements, aligned with much more rapid biological and social processes than the first generation of diagnostic tools could accommodate.

To address the limitations to these foundational definitions, researchers and funding agencies proposed new approaches that recognize the many levels of interacting biological, environmental, and social constructs that are relevant to mental health (Insel et al. 2010). While initiatives like the US National Institute of Mental Health's Research Domain Criteria (RDoC) attempted to address the emerging divide between basic and translational neuroscience and clinical paradigms by redefining health and abnormal mental functioning on the basis of observable behavioral and neurobiological disorders, they also led to tension, as accepted experimental designs and statistical approaches were not well suited to a new dimensional theoretical framework that aimed to integrate interdependent layers of biological and environmental information. Instead, practitioners assumed that the emerging tools of bioinformatics would be implemented as appropriate data sources were generated, without any explicit perspective on how that would be done (Galatzer-Levy et al. 2018b, Insel 2017).

Digital phenotyping: moment-by-moment quantification of individual-level human phenotypes in situ using data from personal digital devices

Passive assessment: measurement of individual functioning through sensors and connected devices

Active assessment: measurement of individual functioning through direct report

DIGITAL MEASUREMENT OF HUMAN BEHAVIOR

Multiple simultaneous and symbiotic advances in the areas of sensor technology, ubiquitous computing, distributed computing, connected devices (e.g., smartphones, wearables, augmented/virtual reality devices), and machine learning have facilitated novel systems of health measurement and intervention (Reddy et al. 2018). Models of human behavior can now be developed and deployed on any number of connected devices to create a digital data source or activate an action directly on the device. In this context, evaluation and treatment delivery can occur on the same timescale (Stroud et al. 2019).

Digital phenotyping, a term coined by one of the authors as the "moment-by-moment quantification of individual-level human phenotypes in situ using data from personal digital devices" (Onnela & Rauch 2016, p. 1691), describes an emerging initiative to understand the distributions of behavioral and physiological activity in real-world settings—a context that is especially relevant for assessment of psychological functioning (Onnela & Rauch 2016). What makes this approach possible is the ubiquity of connected devices, used by a large proportion of the world's population, that can be harnessed as data collection tools. With these tools, researchers can directly measure human behavior and physiological activity using sensor data, referred to as passive assessment, rather than relying on self-report or expert observation, referred to as active assessment. The use of passive assessments embedded in real-world contexts provides an opportunity to bridge the translational gap between laboratory-based constructs that utilize high-fidelity measurement paradigms but lack clinically meaningful population estimates (Hirschtritt & Insel 2018, Insel 2017).

Machine Learning Algorithms

The RDoC introduced conceptual and methodological challenges, as it did not articulate acceptable methods to integrate and interpret the large amounts of novel high-dimensional data that the approach proposed (Lilienfeld 2014). In this section, we review concepts in machine learning that are relevant for addressing these emergent high-dimensional, dynamic data problems. Broadly speaking, machine learning refers to a class of algorithms that aim to learn patterns from data with the goal of making predictions. While several different machine learning methods can be applied in diverse scenarios, they all share basic underlying principles.

Supervised learning: regression and classification. A supervised machine learning model aims to learn a function or rule that maps an input x (i.e., treatment condition, demographics, biomarkers) to an output y (symptom change) across each observation i (subject in the treatment trial) in the data set. The output can be categorical (supervised classification) or continuous (supervised regression) (James et al. 2013, Mohri et al. 2018). If the goal is prediction, then the goal may simply be to identify a model that accurately predicts treatment response in new cases. If the goal is inference, then the goal may be to identify prominent predictors or relationships that could be used to further understand drivers of treatment response.

Supervised learning, specifically regression, allows investigation of how average responses vary across individuals in relation to a set of predictors, such as differential treatment response based on gender, age, or other inputs. The regression curve is defined as the conditional expectation $E(y \mid x)$, that is, the mean value of the response variable y for those observations that have their predictor value equal to x. In the simplest case, such as when studying the relationship between sleep (x) and cognitive performance (y), both x and y are one dimensional. If the problem is low dimensional (as it is here) and we have a very large data set, the task is simple: We find those observation pairs (x_i, y_i) for which x_i is equal (or very close) to the value x, then compute the mean of the corresponding y_i values. However, as we add more predictors, such as age, ethnicity, education, native language, and region, it becomes exponentially more difficult to identify observations that have predictor values equal (or close) to x.

To alleviate this problem, researchers make assumptions about the functional form of $E(y \mid x)$, which leads to parametric regression. Now, instead of having to estimate $E(y \mid x)$ for all values of x, we have to estimate a considerably smaller number of parameters that govern the shape of the regression curve, given its functional form. Point estimates of these parameters, together with their confidence intervals, allow us to infer the values of the parameters and the precision with which they have been estimated from data, respectively. This method also enables predictions of y for new values of x in a manner that propagates our uncertainty about parameters to uncertainty about predictions. For example, if we have two predictors, x_1 and x_2, we can express a linear regression equation as $y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \varepsilon_1$, where β_0, β_1, and β_2 are unknown parameters that need to be estimated from data and ε is an unobserved error. Once we have the parameter estimates, $\hat{\beta}_0$, $\hat{\beta}_1$, and $\hat{\beta}_2$, we can predict the mean response, given x_1 and x_2, as $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2$.

It might seem that including more variables in a model should result in better performance, but this may not be the case. The number of variables (features) included in the model is known as the dimensionality of the problem; the curse of dimensionality refers to the exponential growth in the difficulty of problems as the dimensionality of the problem increases (Luus 2000, pp. 67–80). The fundamental reason for the curse of dimensionality is that in high dimensions (i.e., in high-dimensional Euclidean spaces) all sample points are close to an edge of the sample and, consequently, learning methods need to extrapolate from neighboring sample points rather than interpolate between them (Hastie et al. 2001). From the practical modeling point of view, this

means that adding features that are truly associated with the response should improve model fit but adding noisy features will likely impair fit and exacerbate the risk of overfitting, leading to poor performance outside of training (James et al. 2013).

The assumption about the functional form between the input and the output alleviates the curse of dimensionality considerably, but at the cost of model flexibility. These assumptions, however, enable estimation using relatively small samples, are unlikely to overfit the training data, and may be justifiable on the basis of subject matter considerations. We emphasize that this commonly used class of linear models does not assume a linear relationship between the input and the output. In other words, we do not assume that the model is linear in predictors; instead, we assume that the model is linear in parameters.

To continue with the above example, an updated model might read $y_i = \beta_0 + \beta_1 x_{i1} + \beta x_{i2}^2 + \beta_3 x_{i3} + \varepsilon_1$, which includes the quadratic predictor x_{i2}^2. The updated model no longer has a linear relationship between the input and the output because of the inclusion of the term x_{i2}^2; however, it still sits firmly in the family of linear models because the output is still a linear combination (a weighted average) of the parameters, where the weights are the predictors. This means that even the family of linear models is capable of accommodating various types of nonlinearities between the input and the output. However, if the model is not linear in parameters, such as in $y_i = \beta_0 + \beta_1 \exp(\beta_2 x_i) + \varepsilon_1$, a more general approach is needed.

The counterpoint of parametric regression is nonparametric regression (which applies equally to classification problems), where the regression curve $E(y \mid x)$ is not constrained to follow a specific functional form. Statistics, statistical learning, and machine learning have developed various methods to address nonparametric regression. A common motivation for these methods is that data may show nonlinear patterns with no clear functional forms. These approaches have become more common as data set sizes have increased over time and, in some cases, as the number of predictors has grown. In the 1900s, data sets might have consisted of hundreds or thousands of manually compiled measurements, whereas today tens of millions of observations (or more) may be used to train a model (Goodfellow et al. 2016).

For example, a neural network is a supervised machine learning framework (**Figure 1**) that is commonly utilized in scenarios with a large number of inputs and low a priori knowledge about their relationship to the output. Neural networks consist of simple building blocks, neurons, which are grouped into hidden layers. In a feedforward network, the input x enters from the left and is propagated forward in the network one layer at a time, and the predicted model output $\hat{y}$ is obtained on the right. Neural networks differ in the type of neuron, number of layers, number of neurons per layer, and how the layers are connected. This model specification implicitly codifies the regression function $E(y \mid x)$. If the network has only one neuron, the model may be identical to a logistic regression. Each neuron accepts one or more inputs, calculates a weighted average of them, maps this average through a nonlinear function resembling the sigmoid function familiar from logistic regression, and passes the output to the next layer.

A common and intuitive visual representation is to consider the weights to be attached to the edges of the network. The network structure codifies the architecture of the model. Feedforward networks essentially perform a sequence of logistic regressions on the given network structure. During training, the error estimate (which is a function of observed output and predicted output) leads to an adjustment of the network weights through a process known as back-propagation. This process is repeated until the model converges, which occurs once the minimal error has been reached and subsequent iterations do not lead to further reductions in the error. In both classification and regression settings, the ultimate goal is to learn a function, codified by the specified network structure and the set of learned weights, that provides accurate predictions for new cases, even if the best solution is not interpretable to a human observer.

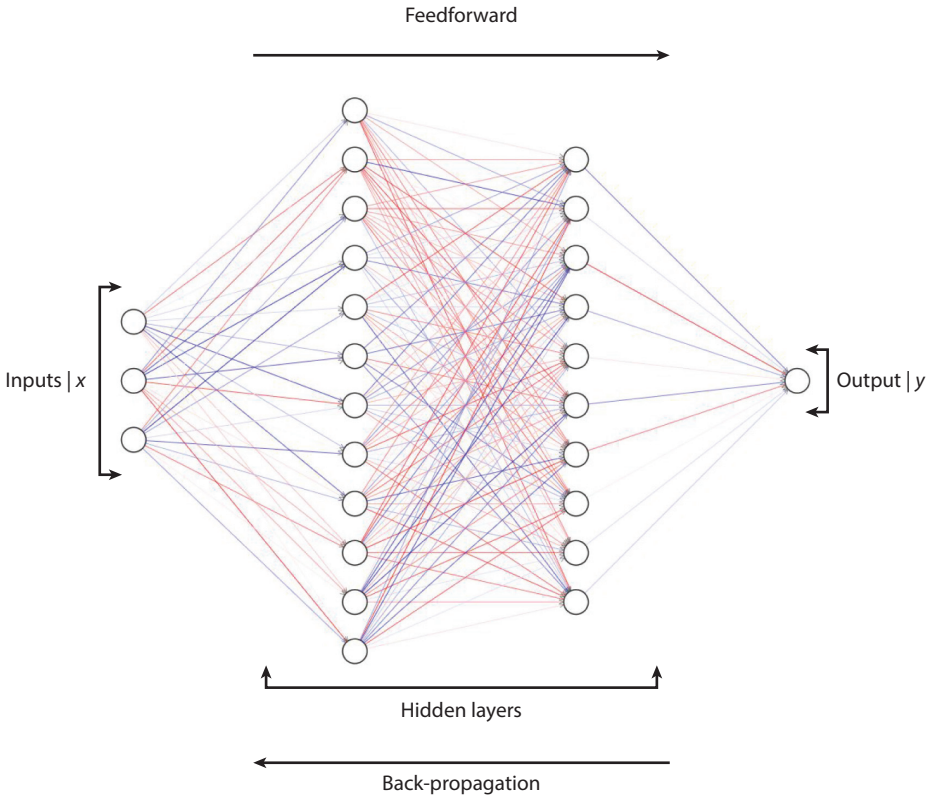

Figure 1

In this synthetic example, three inputs (x) are fed into a fully connected neural network with one hidden layer that has 12 nodes and 1 hidden layer that has 10 nodes, followed by one output layer (y). The connections between nodes are color coded to represent their direction (*blue* denotes positive; *red*, negative) and the strength of the connection (color boldness) as established through back-propagation.

To summarize, a neural network attempts to predict the output y (a class or continuous variable) given the input x with few assumptions about the functional relationship between x and y. In the feedforward step, the input x is fed through the hidden layers, which first forms a weighted average of the inputs and then transforms it through a nonlinear, sigmoid-like function to generate an output y. The error is evaluated by comparing the true observed output y_i with the model-generated predicted output $\widehat{y}_i$. Next, the contribution of each weight to the error is evaluated through a process termed back-propagation, which starts on the right and proceeds to the left, one network layer at a time. The goal of back-propagation is to adjust network weights such that the target y_i and the prediction $\widehat{y}_i$ are closer to one another. This process is repeated to narrow down the combination of weights that best predict the output. In practice, neural networks have several hidden layers, which are flexible enough to model complex relationships and yet are relatively easy to train (see **Figure 1**).

Unsupervised learning: clustering and dimension reduction. The goal of unsupervised learning is to understand relationships between inputs without reference to an outcome (Bernardo & Smith 2001, Mohri et al. 2018). Typically, researchers use unsupervised learning to understand high-dimensional relationships by simplifying them to clusters or lower-dimensional

representations to make sense of complex data sets. Examples of common dimensionality reduction methods are principal components analysis (PCA) and related methods (such as factor analysis), which have long been utilized by psychometricians to empirically identify clusters of symptoms, cognitive functions, personality indicators, or other psychological indicators (Levine & Hunter 1971, Reise et al. 2000). A method closely related to latent variable modeling is latent population clustering (Muthén & Muthén 2000). Methods like latent class analysis are used to identify hidden mixtures of distributions underlying a single observable distribution (Everett 2013) and are commonly applied to identify common trajectories in experimental and clinical data (Galatzer-Levy 2014, Galatzer-Levy et al. 2018a).

Latent variable: a variable that is a mathematically abstracted representation of the relationship between observed variables; for example, depression is a latent variable derived from observed symptom measures

Latent population: a population, such as an economic class, that is best defined by distinct distributions underlying a larger observable distribution

Mixture modeling was first developed to mathematically estimate latent populations from observed data. In 1894, the mathematician Karl Pearson was invited to examine data on the width of Naples crabs. Inconsistent with the hypothesis that a random population phenomenon in nature follows a normal distribution, the distribution across the 1,000 measured crabs was both skewed and kurtotic. Using the method of moments, Pearson analyzed the data by hand and identified two latent populations of crabs, represented by two normal size distributions (Muthén & Muthén 2000). At the time, Pearson hypothesized that there were two genetically distinct populations among the samples, though he could not test that hypothesis empirically.

The general principle of mixture modeling can be extended to longitudinal data to examine change over time. This approach utilizes repeated measures to estimate a set of latent variables that indicate general levels on a particular variable (intercept parameter) and change across measurement occasions (e.g., slope, quadratic parameters). From these variables, latent growth mixture modeling (LGMM) attempts to identify a second-order latent variable (class), which is defined by clusters based on these parameters (Muthén & Asparouhov 2008). As an example, LGMM has been widely used to identify trajectories of response to stress and significant traumatic events to differentiate populations of resilience, recovery, and chronic stress (Galatzer-Levy et al. 2018a).

Semi- and self-supervised machine learning: learning from incomplete data. An exciting and rapidly growing area of machine learning is semisupervised learning. Semisupervised models earn their name from their use of partial or incomplete data to train a model. Such models sample part of the available information, such as 30% of an image or 80% of the words in a book, and attempt to predict the missing parts by using features they derive by analyzing the structure and content of the available data (Floridi & Chiriatti 2020). Such models use imperfect data to generate a lower-dimensional representation of the distribution of features, such as syntax or word usage patterns across sentences, paragraphs, or chapters. This representation is known as an encoder because it encodes information about the object. The model, referred to as a decoder, is then used to generate predictions about the missing information.

The goal of such models is to mimic how humans and animals learn and represent very high-dimensional information and relationships with imperfect and partial information from the outside world, such as images, language, or sequences of actions (Liu et al. 2023, Matsuo et al. 2022). As an example, deepfakes, which are synthetic videos of real people speaking and acting, utilize large corpuses of videos and images to encode high-dimensional information into a lower-dimensional representation that is then applied to generate, or decode, a predicted image under new conditions—for example, the sight and sound of an individual giving a speech that was not used as training data.

Ensemble methods: combined and convoluted models. In real-world scenarios, even simple decisions are highly complex and involve multiple steps. As an example, imagine that a clinical researcher wants to develop a model that classifies eye blinks to measure the frequency of blinking in a video. The measurement of eye blinks cannot be achieved by using only one model; in fact,

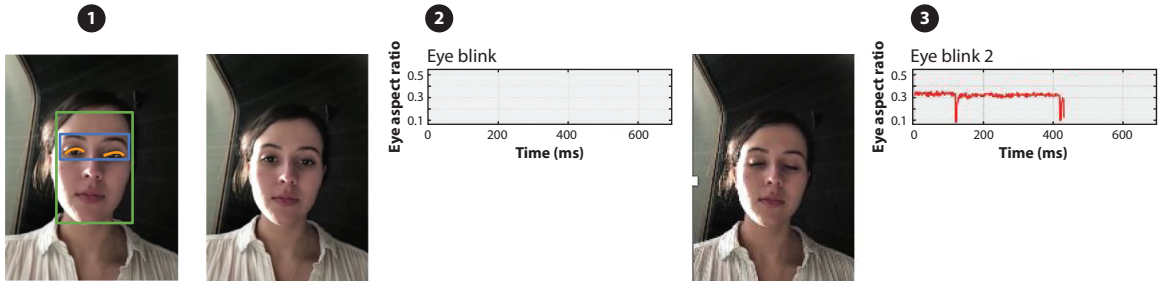

Figure 2

Example of computer vision (CV) measurement of eye blinks. In the eye blink measurement algorithm, (❶) a set of supervised CV classification models are utilized to identify specific features, including the head (*green box*), eyes (*blue box*), and eyelids (*orange ellipse*). Each model is an expert on a specific aspect of the high-dimensional feature space and has been trained on a large number of examples of that specific feature. (❷) Multiple expert models are assembled to determine the ratio of eyelid movement, which is then used as a feature to predict labels of eye blinks. (❸) Eye blinks are registered and counted on the basis of predefined changes in the eyelid-to-eye ratio, providing a count variable and a probability for each eye blink occurrence.

a combination of many models, forming an ensemble, is needed to capture this complex action (**Figure 2**). At a minimum, a model of eye blinks will combine a face identification model, an eye identification model, an eyelid identification model, and an eyelid-open versus eyelid-closed model with a time-series aspect ratio that corresponds to labeled blinking. Each individual model can be conceptualized as an expert in some aspect of the feature space (e.g., eyes open), and the meta-model assembles the various experts in the ensemble to make larger decisions about blinking (Schapire 1990). Additional experts can be utilized to improve model stability, such as models for other facial landmarks that help place the eyes (e.g., nose, ears, lips) or even time-series aspects of the feature space such as lighting changes.

To construct such models, machine learning methods often consist of multiple layers and combinations of models that can include any class of model (unsupervised, supervised, or semisupervised). By combining, or ensembling, multiple weak models that learn a small aspect of the overall task, supervised machine learning models can produce robust, or strong, learners of highly complex tasks (Mohri et al. 2018, Zhou 2021). In the eye blink example, the development of a supervised model may include many constructed features, each of which may have deep pretrained neural network architecture that can be used in an ensemble to achieve the goal of eye blink detection.

How do such models overcome the limitations of existing statistical models to utilize what can often be millions of parameters to characterize high-dimensional relationships in relatively small data sets and provide accurate predictions? To construct such an algorithm, a researcher may employ a special class of neural networks known as convolutional neural networks (CNNs) (LeCun & Bengio 1998). When applied to images, this area of research is known as computer vision (CV). Image and video data consist of three large arrays of numbers representing the red–green–blue (RGB) decomposition of colors for each pixel. Analyzing all possible relationships in very large data sets cannot be achieved with a traditional neural network approach that examines the relationship between every set of inputs and outputs. CNNs overcome this problem in structured data sets, such as images or text, by identifying and combining, or convolving, many simple patterns across the entire image, such as vertical and horizontal edges and distributions of colors. The same process that is applied to single images can be applied to video to detect complex actions. The dynamics of objects themselves become additional layers that are identified, encoded, and decoded (LeCun & Bengio 1998). The addition of dynamic change in a feature is known as a dynamic neural network (DNN); DNNs are commonly utilized in the analysis of audio and video data (Han et al. 2021).

Computer vision (CV): refers to a class of machine learning algorithms that aim to analyze images and video

Convolutional neural network (CNN): a method in machine learning to identify and combine many simple features to analyze, interpret, and predict very high-dimensional data (such as a video)

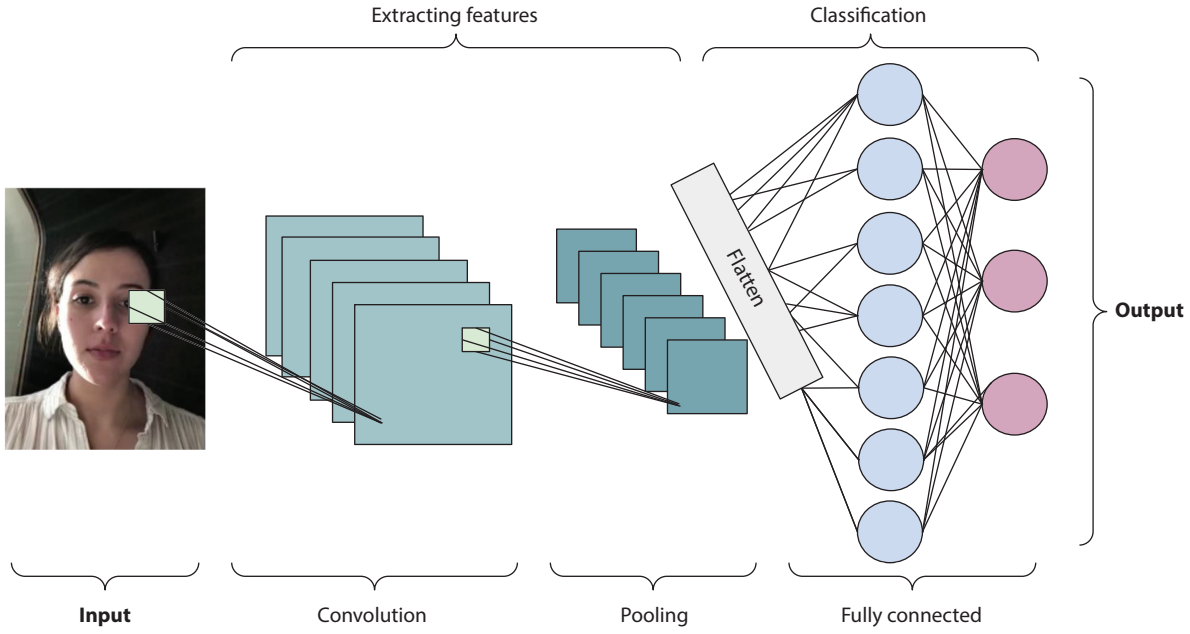

Figure 3

Example of a convolutional neural network to detect eye blinks. First, an image is transformed into three over matrices representing the pixelwise value for red, blue, and green spectra. Next, to extract features from the image, a convolutional layer is applied to small segments of these matrices in sequence to identify simple patterns in each region such as lines and curves. These simple patterns are then pooled and flattened to a smaller dimensional representation. These simple, low-dimensional features are then fed into a fully connected neural network to solve classification or regression problems on the basis of the labels that are provided as an output.

To understand how CV with CNNs works intuitively, consider again the task of training a model to identify eye blinks in a video (**Figure 3**). To do so, the researcher must devise a method to identify a high-dimensional representation of the distribution of features unique to an eye blink. To work in the real world, this algorithm must detect blinking, which involves many characteristics at once across diverse scenarios, including surrounding contexts, angles, lighting conditions, facial characteristics, and perspectives. While the process is highly technical, the goal is to encode information in a manner consistent with how the visual cortex learns and stores representations by integrating simple patterns of shapes and movement to produce robust representations (Boureau et al. 2010). Indeed, differentiating a blink from any other facial characteristic is a trivial task for the human brain. The insight obtained through CNNs is that, while models like neural networks can learn all relationships by connecting all inputs x through a large set of functions, doing so with very high-dimensional data would be impractical. Instead, CNNs aim to learn very simple patterns and intelligently combine them. For example, inputs on the simple shapes of the eye across frames can be used to produce a high-dimensional representation that can then be used to measure a behavioral symptom in a new case at high accuracy.

Connected Digital Devices and Sensor Models

Our ability to measure and model human activity today is fundamentally tied to relatively recent developments in technology. Apart from some high-end music products, vacuum tubes have nearly disappeared and have been universally replaced and superseded by transistors. A transistor is a semiconductor device, famously first demonstrated at Bell Labs on December 23, 1947. Similar to

how atoms combine to form molecules, transistors can be combined in digital electronics to form logical gates whose inputs and outputs are discretized: Low voltage corresponds to zero (Boolean false), and high voltage corresponds to one (Boolean true). Perhaps the most fundamental logical gate is NAND (a negated AND gate), which produces an output that is false only if all its inputs are true. While there are other gates, in principle, every central processing unit empowering any computer can be implemented as a collection of NAND gates, resulting in an astounding number of transistors (>50,000,000,000) that can be combined to perform numerous digital tasks. This unprecedented growth has led to a proliferation of digital devices, especially personal connected digital devices such as computers and smartphones, which are used by an estimated 6.6 billion people worldwide.

Data from connected devices with ever-more-sophisticated sensors that are embedded in these devices can increasingly be used to measure factors such as heart rate, pulse, blood oxygen level, fine and gross motor activity, movement, galvanic skin response, vocal activity, eye movement and dilation, blood sugar, and many other basic vital signs and characteristics. These measures are themselves estimates derived using machine learning models based on data collected with sensors such as photoplethysmography (PPG) sensors, electromyography sensors, gyroscopes, microphones, and cameras.

As the number of connected devices grows, the same information may be collected from distinct platforms. As an example, PPG uses a light-emitting diode to illuminate the skin and then uses a photodiode to measure the amount of light reflected back. This method can be used to measure pulse, heart rate, respiratory rate, blood pressure, and other vital signs (El-Hajj & Kyriacou 2020). Video has also been used to infer pulse and heart rate by measuring displacement across frames in only the red spectrum of RGB matrices (Sikdar et al. 2016). Equivalent models used across sensors have introduced new opportunities. First, if multiple sources provide information, it is possible to develop ensemble models with greater accuracy. Second, these models have introduced the opportunity for device-agnostic models that follow the user across platforms rather than tethering the user to a single device.

Digital Phenotypes

The diverse and growing inputs from connected devices, which provide large quantities of data in real time, have introduced a novel problem in the field of psychology. In contrast to the problems related to the RDC, such as how to access reliable data, psychology is now faced with how to determine the validity of a large quantity of accessible and highly reliable data. Indeed, several sensor-based models are already under study for the collection and analysis of metrics of psychiatric illness and response to treatment (Dorsey et al. 2015, 2017; Shandhi et al. 2021; Van Assche et al. 2022). However, consensus among psychometricians about how to utilize and interpret these measurements for clinically relevant purposes is lacking.

How can new clinical measurements from digital sources be developed and validated? Bridging the gap between current approaches to psychometrics and digital measurement may be more conceptual than technical. In fact, measurement theory is highly consistent with machine learning measurement frameworks—the largest divergence is in the application of models. As an example, psychometricians typically translate the results of a model, such as a screening tool for predicting risk or the factor structure of a scale or test, into a set of hand rules to allow clinicians to calculate latent scores by hand. In contrast, data scientists retain and utilize model-derived weights to simply pass new data through to provide scores or predictions. Retaining this information dramatically increases the predictive accuracy of the same clinical risk screeners (Galatzer-Levy et al. 2014).

The science of psychometrics, more than any other science, has struggled with foundational questions related to the modeling of latent constructs that do not have a ground truth with which

they can be compared (Bollen 2002, McArdle 2009). The lack of ground truth has rapidly emerged as a challenge shared by the fields of neuroscience (Buzsáki 2019) and machine learning (Matsuo et al. 2022). Indeed, there are many ways that digital measurement can reduce the error and burden by providing the technological infrastructure for psychometric concepts to flourish.

As examples, psychometricians utilize factor analysis [a special case of PCA (Bartholomew et al. 2011)] to reduce the dimensionality across items to construct measures of clinical and cognitive performance, and may ensemble many models to derive a higher-order metric. For example, a test or scale utilizes individual items that are summed to subscales such as symptom or cognitive domains, which are then combined into total scores that describe, for example, disorder severity or intelligence quotient (Gootzeit & Markon 2011). Finally, cutoffs are applied that place individuals into distinct classes based on their scores (Weathers et al. 2001). While this process follows the same logic and principles of machine learning models used to find and predict populations, the approach itself introduces hidden errors and loss of information at each step, including the following:

1. Items are assessed on an ordinal rather than an integer scale based on human observation, introducing both subjective bias and non-real-number-ranked values with unknown distance between levels (Annett 2002).
2. Assessments are retrospective, introducing memory biases that are correlated with psychological functioning (Lalande & Bonanno 2011, Wilson et al. 2003).
3. Empirical weights from latent variable models are lost, ignoring the latent probabilities associated with decisions such as cutoffs or class membership. The use of hard cutoffs to define health and illness ignores the facts that there are distributions of functioning within each class and that individuals, especially those close to the cutoff, have a nontrivial probability of belonging to either distribution. For example, individuals in the subsyndromal range for PTSD on traditional measures demonstrate functional impairment equivalent to that of full syndromal individuals (Breslau et al. 2004).

The use of passive digital measurement and machine learning models may reduce error at all levels by providing integer values for behavior and physiology, continuous passive measurement, embedded weights for latent variables, and empirically defined distributions that provide population-specific estimates and a probability for each individual of membership in each of the empirically identified populations.

Example of digital phenotyping I: the application of geolocation to measure daily activity. Reduced social activity and isolation are transdiagnostic symptoms of diverse psychiatric and neurological conditions (Barkus & Badcock 2019, Watson et al. 2017). Social isolation is also easily studied in animal models, leading to a robust body of knowledge of the neurobiological mechanisms, consequences, and potential treatment targets (Mumtaz et al. 2018). Validated self-report approaches to the measurement of isolation are highly burdensome and error prone (Zavaleta et al. 2017). Furthermore, studies of disorders such as depression that are associated with isolating behavior are significantly less likely to provide consistent data (Shrive et al. 2006).

Many connected devices now have embedded GPS sensors that are already utilized for many applications, such as exercise tracking and navigation. Geolocation is accessible to independent developers on many devices through application programming interfaces that can be embedded in the developers' software. Despite the crispness of the data, which provide little error in the estimation of moment-to-moment physical movement, the technology has inherent limitations that require estimation and inference. Because GPS has significant battery requirements, data must be sampled rather than collected continuously. To preserve battery power, a 1-min on cycle

needs to be followed by a much longer off cycle, which may be 10 to 20 min long. This sampling creates a significant missing data problem, creating an additional burden to estimate activity and isolation. Furthermore, GPS, which produces a geolocation by estimating the geometric position based on four satellite signals, is not always robust across environments.

To address this problem, we introduced two missing data imputation methods. In both methods, sequences of latitude–longitude–time triples are aggregated to form two intermediate data types: flights (corresponding to movement) and pauses (corresponding to nonmovement). When there is a gap in the data, the gap is filled by resampling the observed flights and pauses. The challenge, of course, is how to do this well. One of our methods first projects the coordinates to a plane, and then introduces a kernel that assigns resampling weights based on time and location (Barnett & Onnela 2020). Intuitively, the idea is that mobility patterns are tied to time and place, and if we know the time and place of the last observation right before the period of missingness, we can fill the gap based on mobility patterns that occur in that approximate location and time.

This offline method requires input from the beginning of the study up until the current time, and with a hundred thousand or more observations per person per day, it cannot be effectively used in real time. To make the data actionable, we introduced an online variant of the method that, like all online methods, retains a concise summary of all previously seen data, processes the new data in the context of that summary, and then updates the summary based on that day's data (Liu & Onnela 2021). Here, processing refers to imputing the missing GPS trajectories and computing features of interest from the data, such as the amount of time spent at home.

This machine learning method for imputation is termed a Gaussian process model (GP). GPs are a flexible class of models for which any finite-dimensional marginal distribution is assumed to be Gaussian, and can be viewed as potentially infinite-dimensional generalizations of the multivariate Gaussian distribution (Gelman et al. 2013). The variant of this approach used in the second imputation method is technically based on a sparse online variant of a GP. Instead of making assumptions about the functional form of the expectation function $E(y \mid x)$, these models make functional assumptions about the form of the covariance function of the data, which then determines the space of functions the data live in. (For example, so-called radial-basis functions are commonly used choices for the covariance function.) GPs can be considered very flexible, nonparametric extensions of linear regression. One choice for the covariance function leads to the standard linear regression, for which only the first derivative may be nonzero; other choices lead to regression functions that have a potentially infinite number of nonzero derivatives, leading to a very flexible model.

This method was applied to assess daily mobility in people with amyotrophic lateral sclerosis (ALS) during the COVID-19 pandemic (Beukenhorst et al. 2021). (Similar studies are currently in progress in individuals with suicidal ideation and behaviors.) Individuals with ALS generally have low mobility, and the pandemic made them much more homebound than the general population. The median home time prior to the pandemic, estimated from individual-level smartphone GPS data following missing data imputation using our online method, was estimated at 19.4 h [interquartile range (IQR): 15.4–22.0 h]. During the critical early stage of the pandemic, which was accompanied by emergency declarations in several states, the median daily home time for individuals with ALS increased to 23.7 h (IQR: 22.2–24.0 h). This increased time spent at home was also reflected in a reduction in daily mobility, which decreased from 42 km (IQR: 13–83 km) to 3.7 km (IQR: 1.5–10.3 km).

Example of digital phenotyping II: computer vision models of limb tremor. Motor activities such as limb movement, facial activity, and eye movement characteristics represent well-known features of neurological and psychiatric disorders (Walther & Mittal 2022) and have been

cardinal psychiatric symptoms since they were first identified by Kraepelin (1915). However, the evaluation of motor abnormalities such as psychomotor retardation, a symptom of depression and schizophrenia, or limb tremor, a symptom of diverse neurological conditions, requires clinical evaluation in which human raters estimate movement without the use of any external measurement tool. For example, TETRAS (The Essential Tremor Rating Assessment Scale) requires a trained clinician to evaluate a patient's tremor on a scale of one to four across multiple parts of the body (head, face, tongue, voice, upper limb, etc.) by directly observing and ranking motor spastic movement ranges across the body. Upper limb tremor, for example, is evaluated by viewing the patient while extending their arm and ranking the level of tremor as one (barely visible), two (1–3 cm), three (5–10 cm), or four (>20 cm). Scores across parts of the body are summed to produce a total tremor score (Elble 2016).

Ultimately, motor activity is physical action in some dimensional space, making it indexable on a real number scale. Motor activity may be measurable through DNNs that extract and convolve movement layers. Early exploratory work by our research team demonstrated that clinically relevant movement and motor features can be indexed with CNNs and then used to measure symptoms and treatment response (Abbas et al. 2021a,b, 2022; Galatzer-Levy et al. 2021). However, these studies did not measure movement directly but instead produced an approximate metric of movement from the variance of static features such as head position and facial landmark position, making it an unreliable metric of motor activity. To overcome these limitations and produce a clinically meaningful metric, a novel semisupervised CV model, termed a spatiotemporal autoencoder, was constructed to specifically extract a measure of the spatial and temporal dynamics that define movement while ignoring all other information (Zhang et al. 2021).

Instead of relying on flawed measurements as a ground truth, this model (**Figure 4**) aimed to build a model that most closely approximated the physiological parameters of tremor to understand the general distributions of limb movement based on those parameters. To do so, data were experimentally generated by hand tremor using an electrical current in healthy individuals. Semisupervised models were then used to encode movement distributions from videos of limb movement at different levels of manipulation. The distributions thus generated were then used to create a de novo estimate of hand tremor that was then compared with estimates from clinical raters to test convergent validity.

These two examples of digital phenotyping are far from complete developments or validations. However, they both demonstrate convergent principles across divergent approaches, data inputs, and clinically relevant behaviors under investigation. Both sets of models attempt to produce a real-time integer that reflects a human behavior with transdiagnostic relevance. These approaches lower the burden and costs of measurement while increasing precision and scalability. While these examples are closer to proof of concept than scalable technologies, they represent a concrete step forward in the modeling of clinical symptomatology that is being pursued by many researchers in diverse contexts, and will lead to a growing body of knowledge and methods (Dorsey et al. 2017, Liang et al. 2019, Mendes et al. 2022, Onnela 2021, Smith 2018).

THE FUTURE OF DIGITAL PHENOTYPING

The RDC approach aimed to isolate and study main effects while separating out the complexity of psychological symptoms in the real world (Spitzer et al. 1978). Increasingly, psychological functioning is understood to be the product of multiple overlapping biological, social, and environmental systems (Insel et al. 2010). This understanding necessitates research paradigms that are embedded in an individual's real-world context and complex presentation (Nabhan et al. 2019, Sherman et al. 2016). Both data sources and computational approaches are needed to facilitate a new high-dimensional understanding of psychological health and pathology.

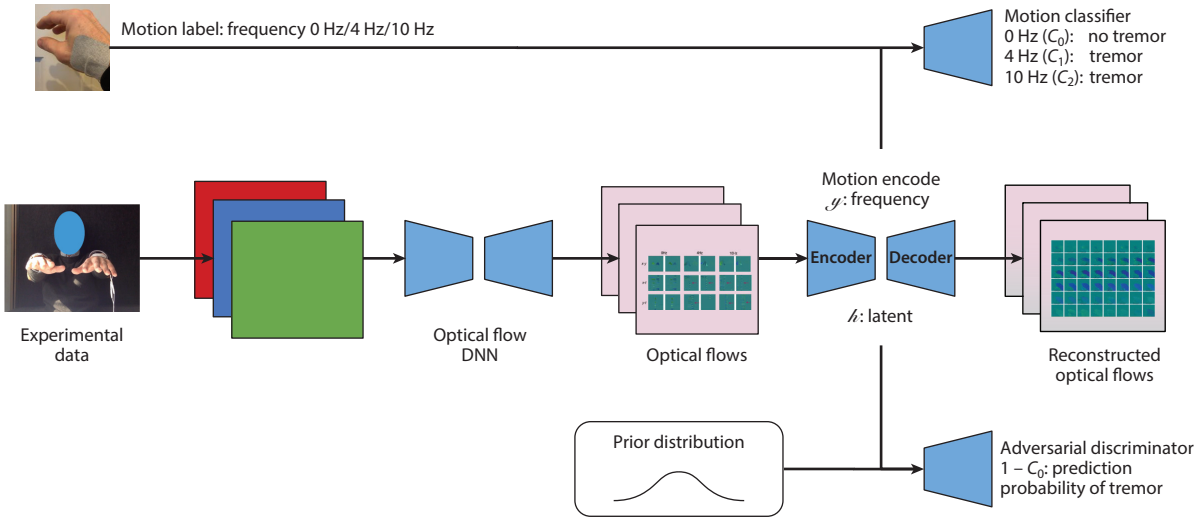

Figure 4

The application of computer vision to measure hand tremor. We utilized the following experimental and modeling steps to predict hand tremor. First, to train models, data from healthy volunteers ($n = 56$) were generated from video recordings using experimental data. Different levels of hand tremor were experimentally induced by passing an electrical current to muscle groups involved in hand tremor using 5 cm × 5 cm electrodes at three preset frequencies administered for 15 s each to match the frequency range of the muscle activity with no tremor (class 0: 0 Hz), moderate tremor (class 1: 4 Hz), and severe tremor (class 2: 10 Hz). Dynamic neural networks were utilized to extract optical flows, with simple features representing movement between frames in a video, including change in height h, width w, number of frames n, and pixelwise displacement in color spectrum from each frame to the next across all frames w. These features were used to produce a hand tremor classification based on the detected frequency across all dimensions (class label $\mathcal{Y}$). Additionally, random samples were drawn from the prior distribution and passed to the decoder subnetwork as latent vectors for image generation; samples generated from the seeds were then pooled with real and augmented samples to boost training data, resulting in a discriminator function that assigned a continuous value representing the probability of tremor presence ($\mathcal{H}$). This score was utilized to test the correlation with expert raters and computer vision estimates of tremor. Abbreviation: DNN, dynamic neural network.

The examples discussed above present new ways in which individual symptoms are being measured using digital connected devices. Though each is itself a complex ensemble of models, ultimately, the output is as fallible an indicator of an underlying clinical construct as any other. Just as Kraepelin observed, populations declare themselves according to the clustering and course of symptoms over time. Increasingly, models that integrate multiple indicators of both human functioning and the environment can be constructed to make predictions about increasingly fine-grained diagnoses and treatment recommendations.

The Future of Passive Digital Measurement in Research

In the context of research, there are clear benefits to a paradigm shift in measurement, including increased sensitivity and frequency, reduced bias, and increased statistical power (Bos et al. 2015, Dodge et al. 2015), accompanied by reduced subject burden and less attrition (Chaix 2018). Passive measurements may be more ecologically valid and better reflect underlying functional paradigms when measured in real-world contexts compared with laboratory-based research settings (Gottman 1993, Ladouce et al. 2016). As an example, the measurement of cognition has long been criticized for recasting broad functions into discrete, isolated units that are fitted to narrow and artificial tasks and do not reflect real-world behavior (Bronfenbrenner 1977, Brunswik 1943, Mayzner & Neisser 1977). Similarly, the meaning of emotional expressions of all kinds is highly

context dependent. Understanding the meaning of these expressions, particularly in differentiating health and pathology, requires the measurement of emotion in the real-world contexts in which they manifest (Aldao & Tull 2015).

Passive measures wrapped around quasi-experimental designs (Campbell & Stanley 2015) that index real-life events and contexts may create the experimental structure necessary to interpret naturalistic data. As an example, the use of unstructured clinical interactions for assessment, such as the type described by First & Spitzer (2003), may return as an important data source as models passively capture important clinical parameters while individuals freely interact. In this context, the algorithm provides the structure to interpret the data, as opposed to the clinician, who is currently forced to create the structure to actively capture clinical information from the patient. Our research has demonstrated that passive measures of facial emotion, rate and pitch of speech, and head and eye movement during open-ended clinical interviews collectively predict PTSD and depression severity, suicide risk, and cognitive functioning (Galatzer-Levy et al. 2021; Schultebraucks et al. 2021b, 2022). Similarly, skin conductance and other physiological markers of autonomic activity among patients in the emergency room are predictive of PTSD course over subsequent months (Hinrichs et al. 2019; Schultebraucks et al. 2020, 2021a).

The Future of Passive Digital Measurement in Clinical Practice

The same tools that have revolutionized digital measurement have concordantly revolutionized digital care. As an example, psychotherapy and psychopharmacology largely moved to digital platforms during the COVID-19 pandemic, allowing them to dramatically increase the reach and scale of treatment but reducing behavioral and physiological observations (Li et al. 2022). Connected devices provide a platform for applications that can take in sensor inputs to instantiate action in real time. Such capabilities provide an opportunity to embed key building blocks of both measurement and intervention into an individual's life in a way that makes treatment targeted and efficient.

Ultimately, the value of any measurement is to predict the correct action. An early example of such an application is NightWare, a smartwatch application that has been approved by the US Food and Drug Administration for the treatment of PTSD. NightWare utilizes commercial sleep algorithms on the Apple Watch to detect when patients diagnosed with PTSD are having sleep disruptions, as measured with green PPG, characteristic of a nightmare. The application then intervenes, using the built-in watch haptics to wake the patient up, resulting in an overall reduction in PTSD over time (Jaklevic 2020, Le Lézard 2020). In this example, both the measurement and the intervention are transdiagnostic and passive, and the ultimate outcome is a broad clinical change.

Such an intervention, grounded in both transdiagnostic symptom measurement and intervention, may find highly diverse clinical applications. Furthermore, both the measurement and the intervention may ultimately become embedded in much more elaborate systems of care that include not only automated actions but also management and feedback to both humans and other automated processes across any number of connected devices. Ultimately, the development of clinically meaningful digital measurements and their implementation in real-world contexts will permit optimized and personalized treatments targeted to the individual's emergent presentation and needs.

CONCLUSION

A new clinical paradigm has emerged from an idealistic vision: accurate, scalable, remote digital measurement that provides the backbone for intelligent, real-time, digital intervention. In this context, evaluation and intervention are closely coupled and dynamically intertwined on the basis of an individual's short- and long-term needs. Data connect clinicians to care that occurs in the

real world. Both clinicians and patients receive actionable feedback based on their response to any given intervention, and over time, care is increasingly fitted to the patient's needs.

This vision is likely articulated in a manner that ultimately exposes the limitations of our current knowledge and the nascent state of digital phenotyping. Despite the limitations associated with prognosticating on the basis of early results, digital phenotyping is a rapidly growing enterprise that is increasingly finding diverse applications in clinical care across distributed platforms. Consistent with prior generations of clinical measurement, assessment matches the problems it intends to solve. At present, both the needs and the technology have outpaced current measurement paradigms, making them ripe for innovation and disruption.

In the context of research, passive digital measurement produces data that better approximate the complex nature of the research subject's life by collecting more granular and multidimensional data with reduced experimental interference. In the clinical context, treatment may become much more personalized, dynamic, and embedded in a manner that complements clinical care by moving assessments and even rote interventions under the management of artificially intelligent machines that learn from open-ended human actions and exchanges among individuals, their environment, and their clinical care providers. The ultimate benefit of this shift is that the artificial apparatus of research and clinical care can fall away from view and provide space for intrinsically human individual differences that drive outcome (McCabe & Priebe 2004).

DISCLOSURE STATEMENT

Isaac R. Galatzer-Levy owns shares in and receives compensation from Meta and Google, where his current and past area of research pertains to sensor-based health measurement. Jukka-Pekka Onnela is a cofounder of a recently established commercial entity that will operate in digital phenotyping.

LITERATURE CITED

Abbas A, Hansen BJ, Koesmahargyo V, Yadav V, Rosenfield PJ, et al. 2022. Facial and vocal markers of schizophrenia measured using remote smartphone assessments: observational study. *JMIR Form. Res.* 6(1):e26276. **https://doi.org/10.2196/26276**

Abbas A, Sauder C, Yadav V, Koesmahargyo V, Aghjayan A, et al. 2021a. Remote digital measurement of facial and vocal markers of major depressive disorder severity and treatment response: a pilot study. *Front. Digit. Health* 3:610006. **https://doi.org/10.3389/fdgth.2021.610006**

Abbas A, Yadav V, Smith E, Ramjas E, Rutter SB, et al. 2021b. Computer vision–based assessment of motor functioning in schizophrenia: use of smartphones for remote measurement of schizophrenia symptomatology. *Digit. Biomark.* 5(1):29–36. **https://doi.org/10.1159/000512383**

Aldao A, Tull MT. 2015. Putting emotion regulation in context. *Curr. Opin. Psychol.* 3:100–7. **https://doi.org/10.1016/j.copsyc.2015.03.022**

Annett J. 2002. Subjective rating scales: science or art? *Ergonomics* 45(14):966–87. **https://doi.org/10.1080/00140130210166951**

Ayer AJ. 1959. *Logical Positivism*. New York: Simon & Schuster

Barkus E, Badcock JC. 2019. A transdiagnostic perspective on social anhedonia. *Front. Psychiatry* 10:216. **https://doi.org/10.3389/fpsyt.2019.00216**

Barnett I, Onnela J-P. 2020. Inferring mobility measures from GPS traces with missing data. *Biostatistics* 21(2):e98–112. **https://doi.org/10.1093/biostatistics/kxy059**

Bartholomew DJ, Knott M, Moustaki I. 2011. *Latent Variable Models and Factor Analysis: A Unified Approach*. New York: Wiley

Bernardo JM, Smith AF. 2001. *Bayesian Theory*. New York: Wiley

Beukenhorst AL, Collins E, Burke KM, Rahman SM, Clapp M, et al. 2021. Smartphone data during the COVID-19 pandemic can quantify behavioral changes in people with ALS. *Muscle Nerve* 63(2):258–62. **https://doi.org/10.1002/mus.27110**

Boesen K, Gøtzsche PC, Ioannidis JPA. 2021. EMA and FDA psychiatric drug trial guidelines: assessment of guideline development and trial design recommendations. *Epidemiol. Psychiatr. Sci.* 30:e35. **https://doi.org/10.1017/S2045796021000147**

Bollen KA. 2002. Latent variables in psychology and the social sciences. *Annu. Rev. Psychol.* 53:605–34. **https://doi.org/10.1146/annurev.psych.53.100901.135239**

Bos FM, Schoevers RA, aan het Rot M. 2015. Experience sampling and ecological momentary assessment studies in psychopharmacology: a systematic review. *Eur. Neuropsychopharmacol.* 25(11):1853–64. **https://doi.org/10.1016/j.euroneuro.2015.08.008**

Boureau YL, Ponce J, LeCun Y. 2010. A theoretical analysis of feature pooling in visual recognition. In *Proceedings of the 27th International Conference on Machine Learning* (*ICML '10*), pp. 111–18. New York: ACM. **https://www.di.ens.fr/willow/pdfs/icml2010b.pdf**

Breslau N, Lucia VC, Davis GC. 2004. Partial PTSD versus full PTSD: an empirical examination of associated impairment. *Psychol. Med.* 34(7):1205–14. **https://doi.org/10.1017/s0033291704002594**

Bronfenbrenner U. 1977. Toward an experimental ecology of human development. *Am. Psychol.* 32(7):513–31. **https://doi.org/10.1037/0003-066X.32.7.513**

Brunswik E. 1943. Organismic achievement and environmental probability. *Psychol. Rev.* 50(3):255–72. **https://doi.org/10.1037/h0060889**

Buzsáki G. 2019. *The Brain from Inside Out*. Oxford, UK: Oxford Univ. Press

Campbell DT, Stanley JC. 2015. *Experimental and Quasi-Experimental Designs for Research*. Cambridge, UK: Ravenio

Carter GC, Cantrell RA, Zarotsky V, Haynes VS, Phillips G, et al. 2012. Comprehensive review of factors implicated in the heterogeneity of response in depression. *Depress. Anxiety* 29(4):340–54. **https://doi.org/10.1002/da.21918**

Casey BJ, Craddock N, Cuthbert BN, Hyman SE, Lee FS, Ressler KJ. 2013. DSM-5 and RDoC: progress in psychiatry research? *Nat. Rev. Neurosci.* 14(11):810–14. **https://doi.org/10.1038/nrn3621**

Chaix B. 2018. Mobile sensing in environmental health and neighborhood research. *Annu. Rev. Public Health* 39:367–84. **https://doi.org/10.1146/annurev-publhealth-040617-013731**

Cusin C, Yang H, Yeung A, Fava M. 2010. Rating scales for depression. In *Handbook of Clinical Rating Scales and Assessment in Psychiatry and Mental Health*, ed. L Baer, MA Blais, pp. 7–35. Totowa, NJ: Humana. **https://doi.org/10.1007/978-1-59745-387-5_2**

Dodge HH, Zhu J, Mattek NC, Austin D, Kornfeld J, Kaye JA. 2015. Use of high-frequency in-home monitoring data may reduce sample sizes needed in clinical trials. *PLOS ONE* 10(9):e0138095. **https://doi.org/10.1371/journal.pone.0138095**

Dorsey ER, Papapetropoulos S, Xiong M, Kieburtz K. 2017. The first frontier: digital biomarkers for neurodegenerative disorders. *Digit. Biomark.* 1(1):6–13. **https://doi.org/10.1159/000477383**

Dorsey ER, Venuto C, Venkataraman V, Harris DA, Kieburtz K. 2015. Novel methods and technologies for 21st-century clinical trials: a review. *JAMA Neurol.* 72(5):582–88. **https://doi.org/10.1001/jamaneurol.2014.4524**

Elble RJ. 2016. The essential tremor rating assessment scale. *J. Neurol. Neuromed.* 1(4):34–38. **https://doi.org/10.29245/2572.942X/2016/4.1038**

El-Hajj C, Kyriacou PA. 2020. A review of machine learning techniques in photoplethysmography for the non-invasive cuff-less measurement of blood pressure. *Biomed. Signal Process. Control* 58:101870. **https://doi.org/10.1016/j.bspc.2020.101870**

Endicott J, Spitzer RL, Fleiss JL. 1975. Mental Status Examination Record (MSER): reliability and validity. *Compr. Psychiatry* 16(3):285–301. **https://doi.org/10.1016/0010-440x(75)90055-3**

Everett B. 2013. *An Introduction to Latent Variable Models*. New York: Springer

First MB, Spitzer RL. 2003. The DSM: not perfect, but better than the alternative. *Psychiatr. Times* 20(4):73–78

Floridi L, Chiriatti M. 2020. GPT-3: its nature, scope, limits, and consequences. *Minds Mach.* 30(4):681–94. **https://doi.org/10.1007/s11023-020-09548-1**

Galatzer-Levy IR. 2014. Empirical characterization of heterogeneous posttraumatic stress responses is necessary to improve the science of posttraumatic stress. *J. Clin. Psychiatry* 75(9):e950–52. **https://doi.org/10.4088/JCP.14com09372**

Galatzer-Levy IR, Abbas A, Ries A, Homan S, Sels L, et al. 2021. Validation of visual and auditory digital markers of suicidality in acutely suicidal psychiatric inpatients: proof-of-concept study. *J. Med. Internet Res.* 23(6):e25199. **https://doi.org/10.2196/25199**

Galatzer-Levy IR, Bryant RA. 2013. 636,120 ways to have posttraumatic stress disorder. *Perspect. Psychol. Sci.* 8(6):651–62. **https://doi.org/10.1177/1745691613504115**

Galatzer-Levy IR, Galatzer-Levy RM. 2007. The revolution in psychiatric diagnosis: problems at the foundations. *Perspect. Biol. Med.* 50(2):161–80. **https://doi.org/10.1353/pbm.2007.0016**

Galatzer-Levy IR, Huang SH, Bonanno GA. 2018a. Trajectories of resilience and dysfunction following potential trauma: a review and statistical evaluation. *Clin. Psychol. Rev.* 63:41–55. **https://doi.org/10.1016/j.cpr.2018.05.008**

Galatzer-Levy IR, Karstoft K-I, Statnikov A, Shalev AY. 2014. Quantitative forecasting of PTSD from early trauma responses: a machine learning application. *J. Psychiatr. Res.* 59:68–76. **https://doi.org/10.1016/j.jpsychires.2014.08.017**

Galatzer-Levy IR, Ruggles K, Chen Z. 2018b. Data science in the research domain criteria era: relevance of machine learning to the study of stress pathology, recovery, and resilience. *Chronic Stress* 2018(2). **https://doi.org/10.1177/2470547017747553**

Gelman A, Carlin JB, Stern HS, Dunson DB, Vehtari A, Rubin DB. 2013. *Bayesian Data Analysis*. Boca Raton, FL: CRC. 3rd ed.

Goodfellow I, Bengio Y, Courville A. 2016. *Deep Learning*. Cambridge, MA: MIT Press

Goodkind M, Eickhoff SB, Oathes DJ, Jiang Y, Chang A, et al. 2015. Identification of a common neurobiological substrate for mental illness. *JAMA Psychiatry* 72(4):305–15. **https://doi.org/10.1001/jamapsychiatry.2014.2206**

Gootzeit J, Markon K. 2011. Factors of PTSD: differential specificity and external correlates. *Clin. Psychol. Rev.* 31(6):993–1003. **https://doi.org/10.1016/j.cpr.2011.06.005**

Gottman JM. 1993. Studying emotion in social interaction. In *Handbook of Emotions*, ed. M Lewis, pp. 475–87. New York: Guilford

Han Y, Huang G, Song S, Yang L, Wang H, Wang Y. 2021. Dynamic neural networks: a survey. *IEEE Trans. Pattern Anal. Mach. Intell.* 44(11):7436–56. **https://doi.org/10.1109/TPAMI.2021.3117837**

Harvey AG, Bryant RA. 1998. The relationship between acute stress disorder and posttraumatic stress disorder: a prospective evaluation of motor vehicle accident survivors. *J. Consult. Clin. Psychol.* 66(3):507–12. **https://doi.org/10.1037//0022-006x.66.3.507**

Hastie T, Friedman J, Tibshirani R. 2001. *The Elements of Statistical Learning*. New York: Springer

Hinrichs R, van Rooij SJ, Michopoulos V, Schultebraucks K, Winters S, et al. 2019. Increased skin conductance response in the immediate aftermath of trauma predicts PTSD risk. *Chronic Stress* 2019(3). **https://doi.org/10.1177/2470547019844441**

Hirschtritt ME, Insel TR. 2018. Digital technologies in psychiatry: present and future. *Focus* 16(3):251–58. **https://doi.org/10.1176/appi.focus.20180001**

Hitchcock PF, Fried EI, Frank MJ. 2022. Computational psychiatry needs time and context. *Annu. Rev. Psychol.* 73:243–70. **https://doi.org/10.1146/annurev-psych-021621-124910**

Hyman SE. 2010. The diagnosis of mental disorders: the problem of reification. *Annu. Rev. Clin. Psychol.* 6:155–79. **https://doi.org/10.1146/annurev.clinpsy.3.022806.091532**

Insel TR. 2017. Digital phenotyping: technology for a new science of behavior. *JAMA* 318(13):1215–16. **https://doi.org/10.1001/jama.2017.11295**

Insel TR, Cuthbert B, Garvey M, Heinssen R, Pine DS, et al. 2010. Research Domain Criteria (RDoC): toward a new classification framework for research on mental disorders. *Am. J. Psychiatry* 167(7):748–51. **https://doi.org/10.1176/appi.ajp.2010.09091379**

Jaklevic MC. 2020. Device approved for adults with nightmare disorder. *JAMA* 324(23):2357. **https://doi.org/10.1001/jama.2020.24228**

James G, Witten D, Hastie T, Tibshirani R. 2013. *An Introduction to Statistical Learning*. New York: Springer

James W. 1890. *The Principles of Psychology*. London: Taylor & Francis

Kessler RC. 1994. The National Comorbidity Survey of the United States. *Int. Rev. Psychiatry* 6(4):365–76. **https://doi.org/10.3109/09540269409023274**

Kessler RC, Chiu WT, Demler O, Merikangas KR, Walters EE. 2005. Prevalence, severity, and comorbidity of 12-month DSM-IV disorders in the National Comorbidity Survey Replication. *Arch. Gen. Psychiatry* 62(6):617–27. **https://doi.org/10.1001/archpsyc.62.6.617**

Kraepelin E. 1915. *Clinical Psychiatry: A Text-Book for Students and Physicians*. New York: Macmillan

Ladouce S, Donaldson DI, Dudchenko PA, Ietswaart M. 2016. Understanding minds in real-world environments: toward a mobile cognition approach. *Front. Hum. Neurosci.* 10:694. **https://doi.org/10.3389/fnhum.2016.00694**

Lalande KM, Bonanno GA. 2011. Retrospective memory bias for the frequency of potentially traumatic events: a prospective study. *Psychol. Trauma* 3(2):165–70. **https://doi.org/10.1037/a0020847**

LeCun Y, Bengio Y. 1998. Convolutional networks for images, speech, and time series. In *The Handbook of Brain Theory and Neural Networks*, ed. MA Arbib, pp. 255–58. Cambridge, MA: MIT Press

LeDoux J. 2012. Rethinking the emotional brain. *Neuron* 73(4):653–76. **https://doi.org/10.1016/j.neuron.2012.02.004**

Le Lézard. 2020. *NightWare receives FDA marketing permission for first and only medical device to reduce sleep disturbances related to PTSD-associated nightmares in adults*. Press Release, Novemb. 20. **https://www.lelezard.com/en/news-19509722.html**

Levine RL, Hunter JE. 1971. Statistical and psychometric inference in principal components analysis. *Multivar. Behav. Res.* 6(1):105–16. **https://doi.org/10.1207/s15327906mbr0601_7**

Li H, Glecia A, Kent-Wilkinson A, Leidl D, Kleib M, Risling T. 2022. Transition of mental health service delivery to telepsychiatry in response to COVID-19: a literature review. *Psychiatr. Q.* 93(1):181–97. **https://doi.org/10.1007/s11126-021-09926-7**

Liang Y, Zheng X, Zeng DD. 2019. A survey on big data–driven digital phenotyping of mental health. *Inf. Fusion* 52:290–307. **https://doi.org/10.1016/j.inffus.2019.04.001**

Liu X, Zhang F, Hou Z, Mian L, Wang Z, et al. 2023. Self-supervised learning: generative or contrastive. *IEEE Trans. Knowl. Data Eng.* 35:857–76

Lilienfeld SO. 2014. The Research Domain Criteria (RDoC): an analysis of methodological and conceptual challenges. *Behav. Res. Ther.* 62:129–39. **https://doi.org/10.1016/j.brat.2014.07.019**

Liu G, Onnela J-P. 2021. Bidirectional imputation of spatial GPS trajectories with missingness using sparse online Gaussian process. *J. Am. Med. Inform. Assoc.* 28(8):1777–84. **https://doi.org/10.1093/jamia/ocab069**

Luus R. 2000. *Iterative Dynamic Programming*. London: Taylor & Francis

Manea L, Gilbody S, McMillan D. 2015. A diagnostic meta-analysis of the Patient Health Questionnaire–9 (PHQ-9) algorithm scoring method as a screen for depression. *Gen. Hosp. Psychiatry* 37(1):67–75. **https://doi.org/10.1016/j.genhosppsych.2014.09.009**

Marshall RD, Spitzer R, Liebowitz MR. 1999. Review and critique of the new DSM-IV diagnosis of acute stress disorder. *Am. J. Psychiatry* 156(11):1677–85. **https://doi.org/10.1176/ajp.156.11.1677**

Matsuo Y, LeCun Y, Sahani M, Precup D, Silver D, et al. 2022. Deep learning, reinforcement learning, and world models. *Neural Netw.* 152:267–75. **https://doi.org/10.1016/j.neunet.2022.03.037**

Mayzner MS, Neisser U. 1977. Cognition and reality. *Am. J. Psychol.* 90(3):541–43. **https://doi.org/10.2307/1421888**

McArdle JJ. 2009. Latent variable modeling of differences and changes with longitudinal data. *Annu. Rev. Psychol.* 60:577–605. **https://doi.org/10.1146/annurev.psych.60.110707.163612**

McCabe R, Priebe S. 2004. The therapeutic relationship in the treatment of severe mental illness: a review of methods and findings. *Int. J. Soc. Psychiatry* 50(2):115–28. **https://doi.org/10.1177/0020764004040959**

Mendes JPM, Moura IR, Van de Ven P, Viana D, Silva FJS, et al. 2022. Sensing apps and public data sets for digital phenotyping of mental health: systematic review. *J. Med. Internet Res.* 24(2):e28735. **https://doi.org/10.2196/28735**

Mohri M, Rostamizadeh A, Talwalkar A. 2018. *Foundations of Machine Learning*. Cambridge, MA: MIT Press. 2nd ed.

Mumtaz F, Khan MI, Zubair M, Dehpour AR. 2018. Neurobiology and consequences of social isolation stress in animal model—a comprehensive review. *Biomed. Pharmacother.* 105:1205–22. **https://doi.org/10.1016/j.biopha.2018.05.086**

Muthén B, Asparouhov T. 2008. Growth mixture modeling: analysis with non-Gaussian random effects. In *Longitudinal Data Analysis*, ed. G Fitzmaurice, M Davidian, G Verbeke, G Molenberghs, pp. 143–66. London: Taylor & Francis

Muthén B, Muthén LK. 2000. Integrating person-centered and variable-centered analyses: growth mixture modeling with latent trajectory classes. *Alcohol. Clin. Exp. Res.* 24(6):882–91

Nabhan C, Klink A, Prasad V. 2019. Real-world evidence—what does it really mean? *JAMA Oncol.* 5(6):781–83. **https://doi.org/10.1001/jamaoncol.2019.0450**

O'Donovan MC. 2015. What have we learned from the Psychiatric Genomics Consortium? *World Psychiatry* 14(3):291–293. **https://doi.org/10.1002/wps.20270**

Onnela J-P. 2021. Opportunities and challenges in the collection and analysis of digital phenotyping data. *Neuropsychopharmacology* 46(1):45–54. **https://doi.org/10.1038/s41386-020-0771-3**

Onnela J-P, Rauch SL. 2016. Harnessing smartphone-based digital phenotyping to enhance behavioral and mental health. *Neuropsychopharmacology* 41(7):1691–96. **https://doi.org/10.1038/npp.2016.7**

Reddy RR, Mamatha C, Reddy RG. 2018. A review on machine learning trends, application and challenges in internet of things. In *2018 International Conference on Advances in Computing, Communications and Informatics* (*ICACCI*), pp. 2389–97. Piscataway, NJ: IEEE

Reise SP, Waller NG, Comrey AL. 2000. Factor analysis and scale revision. *Psychol. Assess.* 12(3):287–97. **https://doi.org/10.1037//1040–3590.12.3.287**

Schapire RE. 1990. The strength of weak learnability. *Mach. Learn.* 5(2):197–227. **https://doi.org/10.1007/BF00116037**

Schultebraucks K, Shalev AY, Michopoulos V, Grudzen CR, Shin S-M, et al. 2020. A validated predictive algorithm of post-traumatic stress course following emergency department admission after a traumatic stressor. *Nat. Med.* 26(7):1084–88. **https://doi.org/10.1038/s41591-020-0951-z**

Schultebraucks K, Sijbrandij M, Galatzer-Levy IR, Mouthaan J, Olff M, van Zuiden M. 2021a. Forecasting individual risk for long-term Posttraumatic Stress Disorder in emergency medical settings using biomedical data: a machine learning multicenter cohort study. *Neurobiol. Stress* 14:100297. **https://doi.org/10.1016/j.ynstr.2021.100297**

Schultebraucks K, Yadav V, Galatzer-Levy IR. 2021b. Utilization of machine learning–based computer vision and voice analysis to derive digital biomarkers of cognitive functioning in trauma survivors. *Digit. Biomark.* 5(1):16–23. **https://doi.org/10.1159/000512394**

Schultebraucks K, Yadav V, Shalev AY, Bonanno GA, Galatzer-Levy IR. 2022. Deep learning–based classification of posttraumatic stress disorder and depression following trauma utilizing visual and auditory markers of arousal and mood. *Psychol. Med.* 52(5):957–67. **https://doi.org/10.1017/S0033291720002718**

Shandhi MMH, Goldsack JC, Ryan K, Bennion A, Kotla AV, et al. 2021. Recent academic research on clinically relevant digital measures: systematic review. *J. Med. Internet Res.* 23(9):e29875. **https://doi.org/10.2196/29875**

Sherman RE, Anderson SA, Dal Pan GJ, Gray GW, Gross T, et al. 2016. Real-world evidence—what is it and what can it tell us? *N. Engl. J. Med.* 375(23):2293–97. **https://doi.org/10.1056/NEJMsb1609216**

Shorter E. 2009. The history of lithium therapy. *Bipolar Disord.* 11(Suppl. 2):4–9. **https://doi.org/10.1111/j.1399-5618.2009.00706.x**

Shrive FM, Stuart H, Quan H, Ghali WA. 2006. Dealing with missing data in a multi-question depression scale: a comparison of imputation methods. *BMC Med. Res. Methodol.* 6:57. **https://doi.org/10.1186/1471-2288-6-57**

Sikdar A, Behera SK, Dogra DP. 2016. Computer-vision-guided human pulse rate estimation: a review. *IEEE Rev. Biomed. Eng.* 9:91–105. **https://doi.org/10.1109/RBME.2016.2551778**

Smith DG. 2018. Digital phenotyping approaches and mobile devices enhance CNS biopharmaceutical research and development. *Neuropsychopharmacology* 43(13):2504–5. **https://doi.org/10.1038/s41386-018-0222-6**

Spitzer RL, Endicott J. 1968. DIAGNO. A computer program for psychiatric diagnosis utilizing the differential diagnostic procedure. *Arch. Gen. Psychiatry* 18(6):746–56. **https://doi.org/10.1001/archpsyc.1968.01740060106013**

Spitzer RL, Endicott J, Robins E. 1975. Clinical criteria for psychiatric diagnosis and DSM-III. *Am. J. Psychiatry* 132(11):1187–92. **https://doi.org/10.1176/ajp.132.11.1187**

Spitzer RL, Endicott J, Robins E. 1978. Research diagnostic criteria: rationale and reliability. *Arch. Gen. Psychiatry* 35(6):773–82. **https://doi.org/10.1001/archpsyc.1978.01770300115013**

Stroud C, Onnela J-P, Manji H. 2019. Harnessing digital technology to predict, diagnose, monitor, and develop treatments for brain disorders. *npj Digit. Med.* 2:44. **https://doi.org/10.1038/s41746-019-0123-z**

Van Assche E, Ramos-Quiroga JA, Pariante CM, Sforzini L, Young AH, et al. 2022. Digital tools for the assessment of pharmacological treatment for depressive disorder: state of the art. *Eur. Neuropsychopharmacol.* 60:100–16. **https://doi.org/10.1016/j.euroneuro.2022.05.007**

Walther S, Mittal VA. 2022. Motor behavior is relevant for understanding mechanism, bolstering prediction, and improving treatment: a transdiagnostic perspective. *Schizophr. Bull.* 48(4):741–48. **https://doi.org/10.1093/schbul/sbac003**

Watson D, Stanton K, Clark LA. 2017. Self-report indicators of negative valence constructs within the Research Domain Criteria (RDoC): a critical review. *J. Affect. Disord.* 216:58–69. **https://doi.org/10.1016/j.jad.2016.09.065**

Weathers FW, Keane TM, Davidson JR. 2001. Clinician-administered PTSD scale: a review of the first ten years of research. *Depress. Anxiety* 13(3):132–56. **https://doi.org/10.1002/da.1029**

Wilson TD, Meyers J, Gilbert DT. 2003. "How happy was I, anyway?" A retrospective impact bias. *Soc. Cogn. N. Y.* 21(6):421–46

Wundt WM, Judd CH. 1902. *Outlines of Psychology*. Leipzig, Ger.: Engelmann

Zavaleta D, Samuel K, Mills CT. 2017. Measures of social isolation. *Soc. Indic. Res.* 131(1):367–91. **https://doi.org/10.1007/s11205-016-1252-2**

Zhang L, Koesmahargyo V, Galatzer-Levy I. 2021. Estimation of clinical tremor using spatio-temporal adversarial autoencoder. In *2020 25th International Conference on Pattern Recognition* (*ICPR*), pp. 8259–66. Piscataway, NJ: IEEE

Zhou Z-H. 2021. Ensemble learning. In *Machine Learning*, pp. 181–210. New York: Springer. **https://doi.org/10.1007/978-981-15-1967-3_8**