Administration and Policy in Mental Health and Mental Health Services Research (2020) 47:795–843
https://doi.org/10.1007/s10488-020-01065-8

ORIGINAL ARTICLE

Improving Mental Health Services: A 50-Year Journey from Randomized Experiments to Artificial Intelligence and Precision Mental Health

Leonard Bickman[1]

Published online: 26 July 2020

Abstract

This conceptual paper describes the current state of mental health services, identifies critical problems, and suggests how to solve them. I focus on the potential contributions of artificial intelligence and precision mental health to improving mental health services. Toward that end, I draw upon my own research, which has changed over the last half century, to highlight the need to transform the way we conduct mental health services research. I identify exemplars from the emerging literature on artificial intelligence and precision approaches to treatment in which there is an attempt to personalize or fit the treatment to the client in order to produce more effective interventions.

Keywords Mental health services · Artificial intelligence · Machine learning · Precision mental health · Randomized clinical trials (RCTs) · Precision medicine

"AI will bring many wonders. It may also destabilize everything from nuclear détente to human friendships. We need to think much harder about how to adapt... it is changing human knowledge, perception and reality—and, in so doing, changing the course of human history. We seek to understand it and its consequences and encourage others across the disciplines to do the same."

—Kissinger et al. 2019, pp. 24–26

“AI is one of the most profound things we’re working on as humanity. It’s more profound than fire or electricity.”

—Alphabet CEO Sundar Pichai, in Thompson and Bodoni 2020.

In 1963, I was writing my first graduate paper at Columbia University on curing schizophrenia using Sarnoff Mednick’s learning theory. I was not very modest even as a first-year graduate student! But I was puzzled as to how to develop and evaluate a cure. Then, as now, the predominant research design was the randomized experiment or randomized clinical trial (RCT). It was clear that simply describing, let alone manipulating, the relevant characteristics of this one disorder and promising treatments would require hundreds of variables. Developing an effective treatment would take what seemed to me an incalculable number of randomized trials. How could we complete all the randomized experiments needed? How many different outcomes should we measure? How could we learn to improve treatment? How should we consider individual differences in these group comparisons? I am sure I was not insightful enough to think of all these questions back then, but I know I felt frustrated and stymied by our methodological approach to answering these questions. But I had to finish the paper, so I relegated these and similar questions to the list of universal imponderables such as why I exist. In fact, I became a committed experimentalist, and I dealt with the limitations of experiments by recognizing their restrictions and abiding by the

This article is part of a special issue, *Festschrift for Leonard Bickman: The Future of Children’s Mental Health Services*.

✉ Leonard Bickman
lbickman@fiu.edu

[1] Center for Children and Families; Psychology, Academic Health Center 1, Florida International University, 11200 Southwest 8th Street, Room 140, Miami, FL 33199, USA

principle “For determining causality, in many but not all circumstances, the randomized design is the worst form of design except all the others that have been tried[1]” (Bickman and Reich 2014, pp. 104–105).

For the much of my career, I was a committed proponent of the RCT as the best approach to understanding causal relationships (Bickman 2006). However, as some of my writing indicates, it was a commitment with reservations. I did not see a plausible alternative or complement to RCTs until recently, when I began to read about artificial intelligence (AI) and precision medicine in 2013. The potential solution to my quandary did not crystallize until 2016, when I collaborated with Aaron Lyons and Miranda Wolpert on a paper on what we called “precision mental health” (Bickman et al. 2016). With the development of AI and its application in precision medicine, I now believe that AI is another approach that we may be able to use to understand, predict, and influence human behavior. While not necessarily a substitute for RCTs in efforts to improve mental health services, I believe that AI provides an exciting alternative to RCTs or an adjunct to them. While I use *precision medicine* and *precision mental health* interchangeably, I will differentiate them later in this paper.

Toward that end, I focus much of this paper on the role of AI and precision medicine as a critical movement in the field with great potential to inform the next generation of research. Before proposing such solutions, I first describe the challenges currently faced by mental health services, using examples drawn almost entirely from studies of children and youth, the area in which I have conducted most of my research. I describe five principal causes of this failure, which I attribute primarily, but not solely, to methodological limitations of RCTs. Lastly, I make the case for why I think AI and the parallel movement of precision medicine embody approaches that are needed to augment, but probably not replace, our current research and development efforts in the field of mental health services. I then discuss how AI and precision mental health can help inform the path forward, with a focus on similar problems manifested in mental health services for adults. These problems, I believe, make it clear that we need to consider alternatives to our predominant research approach to improving services. Importantly, most of the research on AI and precision medicine I cite deals with adults, as there is little research in this area on children and youth. I am assuming that we can generalize from one literature to the other, but I anticipate that there many exceptions to this assumption.

Why I am Dissatisfied With the Current State of Mental Health Services

The impetus for change in mental health services is motivated by my dissatisfaction with the status quo of services for children and youth. I do not believe we should be satisfied with our current services. I briefly review three core problems that support my contention that we should not be content with the current services.

Services are Not Sufficiently Accessible

According to some estimates, more than half (56.4%) of adults with a mental illness receive no treatment (Mental Health in America 2018). Less than half of adolescents with psychiatric disorders receive any kind of treatment (Costello et al. 2014). Over 60% of youth with major depression do not receive any mental health treatment (Mental Health in America 2018). Several other relevant facts when it comes to youth illustrate the problem of their access to services. Hodgkinson et al. (2017) have documented that less than 15% of children in poverty receive needed services. These authors also showed that there is less access to services for minorities and rural families. When it comes to the educational system, Mental Health in America (2018) estimated that less than 1% of students have an Individual Education Plan (IEP), which students need to access school-supported services, even though studies have shown that a much larger percentage of students need those services. Access is even more severely limited in in low- and middle-income countries (Esponda et al. 2020).

Evidence-Based Services are Not Well Implemented

Very few clients receive effective evidence-based quality mental health services that have been shown to be effective in laboratory-based research (Garland et al. 2010; Gyani et al. 2014). Moreover, research shows that even when they do receive care that is labeled evidence-based, it is not implemented with sufficient fidelity to be considered evidence-based (Park et al. 2015). No matter how effective evidence-based treatments are in the laboratory, it is very clear that they lose much of their effectiveness when implemented in the real world (Weisz et al. 2006, 2013).

Services are Not Sufficiently Effective

Research reviews demonstrate that services that are typically provided outside the laboratory lack substantial evidence of effectiveness. There are two factors that account for this lack of effectiveness. As noted above, evidence-based services are usually not implemented with sufficient

[1] This is a paraphrasing of Winston Churchill’s famous quotation about democracy: “Democracy is the worst form of Government except all those other forms that have been tried from time to time” (Parliament Bill, HC Deb 11, November 1947, Vol. 444, cc203-32.

fidelity to replicate the effectiveness found in the laboratory. More fundamentally, it is argued here that even evidence-based services may not be sufficiently effective as currently conceptualized. A review of 23 published studies on school-based health centers found that while these services increased access, the review could not determine whether services were effective because the research was of such poor quality (Bains and Diallo 2016). A meta-analysis of 43 studies of mental health interventions implemented by school personnel found small to medium effect sizes, but only 2% of the services were provided by school counselors or mental health workers (Sanchez et al. 2018). A Cochrane Review concluded, "We do not know whether psychological therapy, antidepressant medication or a combination of the two is most effective to treat depressive disorders in children and adolescents" (Cox et al. 2014, p. 3). Another meta-analysis of 24 studies on school-based interventions delivered by teachers showed a small effect for internalizing behaviors but no effect on externalizing ones (Franklin et al. 2017a). Similarly, a meta-analysis of 74 meta-analyses of universal prevention programs targeting school-age youth showed a great deal of variability with effect sizes from 0 to 0.5 standard deviations depending on type of program and targeted outcome (Tanner-Smith et al. 2018). A review of 32 RCTs found no compelling evidence to support any one psychosocial treatment over another for people with serious mental illnesses (Hunt et al. 2013). A systematic review and meta-analysis of 19 conduct disorder interventions concluded that they have a small positive effect, but there was no evidence of any differential effectiveness by type of treatment (Bakker et al. 2017). Fonagy and Allison (2017) conclude, "The demand for a reboot of psychological therapies is unequivocal simply because of the disappointing lack of progress in the outcomes achieved by the best evidence-based interventions" (p. 978).

Probably the most discouraging evidence was identified by Weisz et al. (2019) on the basis of a review of 453 RCTs over a 50-year period. They found that the mean effect size for treatment did not improve significantly for anxiety and ADHD and decreased significantly for depression and conduct problems. The authors conclude:

> In sum, there were strikingly few exceptions to the general pattern that treatment effects were either unchanged or declining across the decades for each of the target problems. One possible implication is that the research strategy used over the past 5 decades, the treatment approaches investigated, or both, may not be ideal for generating incremental benefit over time. (p. 17)

There is a need—indeed, an urgent need—to change course, because our traditional approaches to services appear not to be working. However, we might be expecting too much from therapy. In an innovative approach to examining the effectiveness of psychotherapy for youth, Jones et al. (2019) subjected 502 RCTs to a mathematical simulation model that estimated that even if therapy was perfectly implemented, the effect size would be a modest 0.83. They concluded that improving the quality of existing psychotherapy will not result in much better outcomes. They also noted that AI may help us understand why some therapies are more effective than others. They suggested that the impact of therapy is limited because a plethora of other factors influence mental health, especially given that therapy typically lasts only one hour a week out of 110 + waking hours. They also indicated that other factors that have not been included in typical therapies, such as individualizing or personalizing treatment, may increase the effectiveness of treatment.

I am not alone in signaling concern about the state of mental health services. For example, other respected scholars in children's services research have also raised concerns about the quality and effectiveness of children's services. Weisz and his colleagues (Marchette and Weisz 2017; Ng and Weisz 2016) described several factors that contribute to the problems identified above. These included a mismatch between empirically supported treatments and mental health care in the real world, the lack of personalized interventions, and the absence of transdiagnostic treatment approaches. It is important to acknowledge the pioneering work of Sales and her colleagues, who identified the need and tested approaches to individualizing assessment and monitoring clients (Alves et al. 2013, 2015; Elliott et al. 2016; Sales and Alves 2012, 2016; Sales et al. 2007, 2014). We need not only to appreciate the relevance of this work but also to integrate it with new artificial intelligence approaches described later in this paper.

I am not concluding from such evidence that all mental health services are ineffective. This brief summary of the state of our services can be perceived in terms of a glass half full or half empty. In other words, there is good evidence that some services are effective under particular, but yet unspecified, conditions. However, I do not believe that the level of effectiveness is sufficient. Moreover, we are not getting better at improving service effectiveness by following our traditional approach to program development, implementation, research, and program evaluation. While it is unlikely that the social and behavioral sciences will experience a major breakthrough in discovering how to "cure" mental illness, similar to those often found in the physical or biological sciences, I am arguing in this paper that we must increase our research efforts using alternative approaches to produce more effective services. A large part of this paper, therefore, is devoted to exploring what has been also called a precision approach to treatment in which there is an attempt to personalize treatment or fit treatment to the client in order to produce more effective interventions.

The Large Investment in Systems of Care Has Distracted Us From Concerns About the Effectiveness of Care

In some of my earliest work in mental health, I identified the field's focus on system-level factors rather than on treatment effectiveness as one cause of the problems with mental health services. The most popular and well-funded approach to mental health services in the 1960s and 1970s, which continues even today, is called *a system or continuum of care* (Bickman 1997, 1999; Bickman et al. 1997b; Bryant and Bickman 1996). This approach correctly recognized the problems with the practice of providing services that were limited to outpatient and hospitalization only, which was very common at that time. Moreover, these traditional services did not recognize the importance of the role played by youth and families in the delivery of mental services. To remedy these important problems, advocates for children's mental health conceptualized that a system of care was needed, in which a key ingredient was a managed continuum of care with different levels or intensiveness of services to better meet the needs of children and youth (Stroul and Friedman 1986). This continuum of care is a key component of a system of care. However, I believe that in actuality, these different levels of care simply represent different locations of treatment and restrictiveness (e.g., inpatient vs. outpatient care) and did not necessarily reflect a gradation of intensity of treatment.

A system of care is not a specific type of program, but an approach or philosophy that combines a wide range of services and supports for these services with a set of guiding principles and core values. Services and supports are supposed to be provided within these core values, which include the importance of services that are community-based, family-focused, youth-oriented, in the least restrictive environment, and culturally and linguistically proficient. System-level interventions focus on access and coordination of services and organizations and not on the effectiveness of the treatments that are provided. It appeared that the advocates of systems of care assumed that services were effective and that what was needed was to organize them better at the systems level. Although proponents of systems of care indicated that they highly valued individualized treatment, especially in what were called wraparound services, there was no distinct and systematic way that individualization was operationalized or evaluated. Moreover, there was not sufficient evidence that supported the assumption that wraparound services produced better clinical outcomes (Bickman et al. 2003; Stambaugh et al. 2007). A key component of the system is providing different levels of care that include hospitalization, group homes, and outpatient services, but there is little evidence that clinicians can reliably assign children to what they consider the appropriate level of care (Bickman et al. 1997a).

The Fort Bragg Study

My earliest effort in mental health services research was based on a chance encounter that led to the largest study ever conducted in the field of child and youth mental health services. I was asked by a friend to see if I could help a person whom I did not know to plan an evaluation of a new way to deliver services. This led to a project that cost about $100 million to implement and evaluate. We evaluated a new system of care that was being implemented at Fort Bragg, a major U.S. Army post in North Carolina. We used a quasi-experimental design because the Army would not allow us to conduct a RCT; however, we were able to control for many variables by using two similar army posts as controls (Bickman 1995; Bickman et al. 1995). The availability of sufficient resources allowed me to measure aspects of the program that were not commonly measured at that time, such as cost and family empowerment. With additional funding that I received from a competitive grant from the National Institute of Mental Health (NIMH) and additional follow-up funding from the Army, we were able to do a cost-effectiveness analysis (Foster and Bickman 2000), measure family outcomes (Heflinger and Bickman 1996), and develop a new battery of mental health symptoms and functioning (Bickman and Athay 2012). In addition, we competed successfully for an additional NIMH grant to evaluate another system of care in a civilian population using a RCT (Bickman et al. 1997a, b) and a study of a wraparound services that was methodologically limited because of sponsor restrictions (Bickman et al. 2003).

I concluded from this massive and concentrated effort that systems of care (including the continuum of care) were able to influence system-level variables, such as access, cost, and coordination, but that there was not sufficient evidence to support the conclusion that it produced better mental health outcomes for children or families or that it reduced costs per client (Bickman et al. 2000). This conclusion was not accepted by the advocates for systems of care or the mental health provider community more generally. Moreover, I became *persona non grata* among the proponents of systems of care. While the methodologists who were asked to critique on the Fort Bragg study saw it as an important but not flawless study (e.g., Sechrest and Walsh 1997; Weisz et al. 1997) that should lead to new research (Hoagwood 1997), most advocates thought it to be a well-done evaluation but of very limited generalizability (Behar 1997).

The System of Care Approach Today

It is important to note that the system of care approach, almost 30 years later, remains the major child and youth program funded by the Substance Abuse and Mental Health Services Administration's (SAMHSA) Center for Mental Health Services (CMHS) to the tune of about a billion dollars in funding since the system of care program's inception in 1993. There have been many evaluations funded as part of the SAMHSA program that show some positive results (e.g., Holden et al. 2001), but, in my opinion, they are methodologically weak and, in some cases, not clearly independent. Systems of care are still considered by SAMHSA's Center for Mental Health Services to be the premier child and adolescent program worthy of widespread diffusion and funding (Substance Abuse and Mental Health Services Administration 2019), regardless of what I believe is the weak scientific support. This large investment of capital should be considered a significant opportunity cost that has siphoned off funds and attention from more basic concerns such as effectiveness of services. Sadly, based on my unsuccessful efforts to encourage change as a member of the CMHS National Advisory Council (2019–2023), I am not optimistic that there will be any modification of support for this program or shift of funding to more critical issues that are identified in this paper.

In the following section, I consider some of the problems that have contributed to the current status of mental health services.

Five Problems that Contribute to Poor Services

My assessment of current services led me to categorize the previously described deficiencies into the five following related problem groups.

The Problem of Diagnoses Muddle

The problems with the validity of diagnoses have existed for as long as we have had systems of diagnoses. While a diagnosis provides some basis for tying treatment to individual case characteristics, its major contribution is providing a payment system for reimbursement for services. Research has shown that external factors such as insurance influence the diagnosis given, and the diagnosis located in electronic health records is influenced by commercial interests (Perkins et al. 2018; Taitsman et al. 2020). Other studies have demonstrated that the diagnosis of depression alone is not sufficient for treatment selection; additional information is required (Iniesta et al. 2016). Moreover, others have shown that diagnostic categories overlap and are not mutually exclusive (Bickman et al. 2012c). In practice, medication is prescribed according to symptoms and not diagnosis (Waszczuk et al. 2017).

In their thematic analysis of selected chapters of the *Diagnostic and Statistical Manual of Mental Disorders* (*DSM–5)*, Allsopp et al. (2019) examined the heterogeneous nature of categories within the *DSM-5*. They showed how this heterogeneity is expressed across diagnostic criteria, and explained its consequences for clinicians, clients, and the diagnostic model. The authors concluded that "a pragmatic approach to psychiatric assessment, allowing for recognition of individual experience, may therefore be a more effective way of understanding distress than maintaining commitment to a disingenuous categorical system" (p. 15). Moreover, in an interview, Allsop stated:

> Although diagnostic labels create the illusion of an explanation, they are scientifically meaningless and can create stigma and prejudice. I hope these findings will encourage mental health professionals to think beyond diagnoses and consider other explanations of mental distress, such as trauma and other adverse life experiences. (*Neuroscience News* 2019, para. 6)

Finally, a putative solution to this muddle is NIMH's Research Domain Criteria Initiative (RDoC) diagnostic guide. RDoC is not designed to be a replacement of current systems but serves as a research tool for guiding research on mental disorders systems. However, it has been criticized on several grounds. For example, Heckers (2015) states, "It is not clear how the new domains of the RDoC matrix map on to the current dimensions of psychopathology" (p. 1165). Moreover, there is limited evidence that RDoC has actually improved the development of treatments for children (e.g., Clarkson et al. 2019). As I will discuss later in the paper, Rush and Ibrahim (2018), in their critical review of psychiatric diagnosis, predicted that AI, especially artificial neural networks, will change the nature of diagnosis to support precision medicine.

The Problem of Poorly Designed Measures

If measures are going to be used in real world practice, then in addition to the classic and modern psychometric validity criteria, it must be possible to use measures sufficiently often to provide a fine-grained picture of change. If measures are used frequently, then they must be short so as not to take up clinical time (Riemer et al. 2012). Moreover, since there is a low correlation among different respondents (De Los Reyes and Ohannessian 2016), we need measures and data from different respondents including parents, clinicians, clients, and others (e.g., teachers). However, we are still lacking a systematic methodology for managing these different perspectives.

Since we are still unsure which constructs are important to measure, we need measures of several different constructs in order to pinpoint which ones we should administer on a regular basis. In addition to outcome measures, we need valid and reliable indicators of mediators and processes to test theories of treatment as well as to indicate short-term outcomes. We need measures that are sensitive to change to be valid measures of improvement. We need new types of measures that are more contextual, that occur outside of therapy sessions, and that are not just standardized questionnaires. We lack good measures of fidelity of implementation that capture in an efficient manner what clinicians actually do in therapy sessions. This information is required to provide critical feedback to clinicians. We also lack biomarkers of mental illness that can be used to develop and evaluate treatments that are often found in physical illnesses.

This is a long and incomplete list of needs and meeting them will be difficult to accomplish without a concerted effort. There are some resources at the National Institutes of Health that are focused on measure development, such as Patient-Reported Outcomes Measurement System Information (PROMIS) (https://www.healthmeasures.net/explore-measurement-systems/promis), but this program does not focus on mental health. Thus, we depend upon the slow and uncoordinated piecemeal efforts of individual researchers to somehow fit measure development into their career paths. I know this intimately because when I started to be engaged with children's mental health services research, I found that the measures in use were too long, too expensive, and far from agile. This dissatisfaction led me down a long path to the development of a battery of measures called the Peabody Treatment Progress Battery (Bickman and Athay 2012; Riemer et al. 2012). This battery of 12 brief measures was developed as part of ongoing research grants and not with any specific external support.

The Problem of the Primacy of RCTs

For over a half century, I have been a committed experimentalist. I still am a big fan of experiments for some purposes (Bickman 2006). The first independent study I conducted was my honors thesis at City College of New York in 1966. My professor was a parapsychologist and personality psychologist, so the subject of my thesis was extrasensory perception (ESP). My honors advisor had developed a theory of ESP that predicted that those who were positive about ESP, whom she called sheep, would be better at ESP than the people who rejected ESP, whom she called goats (Schmeidler 1952). Although I did not realize it at the time, my experimentalist or action orientation was not satisfied with correlational findings that were the core of the personality approach. I designed an experiment in which I randomly assigned college students to hear a scripted talk from me supporting or debunking ESP. I found very powerful results. The experimental manipulation changed people's perspective on the efficacy of ESP, but I found no effect on actual ESP scores. It was not until I finished my master's degree in experimental psychopathology at Columbia University that I realized that I wanted to be an experimental social psychologist, and I became a graduate student at the City University of New York. However, I did not accept the predominant approach of social psychologists, which was laboratory experimentation. I was convinced that research needed to take place in the real world. Although my dissertation was a laboratory study of helping behavior in an emergency (Bickman 1972), it was the last lab study I did that was not also paired with a field experiment (e.g. Bickman and Rosenbaum 1977). One of my first published research studies as a graduate student was a widely cited field experiment (RCT) that examined compliance to men in different uniforms in everyday settings (Bickman 1974a, b).

The first book I coedited, as a graduate student, was titled *Beyond the Laboratory: Field Research in Social Psychology* and was composed primarily of field experiments (Bickman and Henchy 1972). Almost all my early work as a social psychologist consisted of field experiments (Riemer and Bickman 2011). I strongly supported the primacy of randomized designs in several textbooks I coauthored or coedited (Alasuutari et al. 2008; Bickman and Rog 2009; Bickman and Rog 2016; Hedrick et al. 1993). While the Fort Bragg study I described above was a quasi-experiment (Bickman 1996), I was not happy that the funding agency, the U.S. Army, did not permit me to use a RCT for evaluating an important policy issue. As I was truly committed to using a RCT to evaluate systems of care, I followed up this study with a conceptual replication in a civilian community using a RCT (Bickman et al. 1997b) that was funded by a NIMH grant. While I have valued the RCT and continue to do so, I have come to the conclusion that our experimental methods were developed for simpler problems. Mental health research is more like weather forecasting with thousands of variables rather than like traditional experimentation, which is based on a century-old model for evaluating agricultural experiments with only a few variables (Hall 2007). We need alternatives to the traditional way of doing research, service development, and service delivery that recognize the complexity of disorders, heterogeneity of clients, and varied contexts of mental health services. The oversimplification of RCTs has produced a blunt tool that has not served us well for swiftly improving our services. This is not to say that there has been no change in the last 75 years. For example, the Institute of Education Sciences, a more recent player the field of children's behavioral and mental health outcomes research, has released an informative monograph on the use of adaptive randomized trials that does demonstrate flexibility in describing how RCTs can be implemented in innovative ways (Nahum-Shani and Almirall 2019).

The concerns about RCTs are also apparent in other fields. For example, a special issue of *Social Science and Medicine* focused on the limitations of RCTs (Deaton and Cartwright 2018). The contributors to this incisive issue indicated that a RCT does not in practice equalize treatment and control groups. RCTs do not deliver precise estimates of average treatment effects (ATEs) because a RCT is typically just one trial, and precision depends on numerous trials. There is also an external validity problem; that is, it is difficult to generalize from RCTs, especially those done in university laboratory settings. Context is critical and theory confirmation/disconfirmation is important, for without generalizability, the findings are difficult to apply in the real world (Bickman et al. 2015).

Scaling up from a rigorous RCT to a community-based treatment is now recognized as a significant problem in the relatively new fields of translational research and implementation sciences. In addition to scaling up, there is a major issue in scaling down to the individual client level. Stratification and theory help, but they are still at the group level. The classic inferential approach also has problems with replication, clinical meaningfulness, accurate application to individuals, and *p*-value testing (Dwyer et al. 2018).

The primary clinical problem with RCTs is the emphasis on average treatment effects (ATEs) versus individual prediction**.** RCTs emphasize postdiction, and ATEs lead to necessary oversimplification and a focus on group differences and not individuals. Subramanian et al. (2018) gave two examples of the fallacy of averages: The first was a 1942 study to describe the "ideal woman," where they measured nine body dimensions and then averaged each one. A contest to identity the "average woman" got 4000 responses, but not a single woman matched the averages on all nine variables. In a second example, the U.S. Air Force in 1950 measured 400 pilots on 140 body dimensions to determine appropriate specifications for a cockpit. Not a single pilot matched the averages on even as few as 10 dimensions, even when their measurements fell within 30% of the mean value. As these examples show, the problem with using averages has been known for a long time, but we have tended to ignore this problem. We are disappointed when clinicians do not use our research findings when in fact our findings may not be very useful for clinicians because clinicians deal with individual clients and not some hypothetical average client. We can obtain significant differences in averages between groups, but the persons who actually benefit from therapy will vary widely to the extent to which they respond to the recommended treatments. Thus, the usefulness of our results depends in part on the heterogeneity of the clients and the variability of the findings.

The privileging of RCTs also came with additional baggage. Instead of trying to use generalizable samples of participants, the methodology favored the reduction of heterogeneity as a way to increase the probability of finding statistically significant results. This often resulted in the exclusion from studies of whole groups of people, such as women, children, people of color, and persons with more than one diagnosis. While discussions often included an acknowledgment of this limitation, little was done about these artificial limitations until inclusion of certain groups was required by federal funding agencies (National Institutes of Health, Central Resource for Grants and Funding Information 2001).

The limitations of RCTs are not a secret, but we tend to ignore these limitations (Kent et al. 2018). One attempt to solve the difficulty of translating average effect sizes by RCTs to individualize predictions is called *reference class forecasting*. Here, the investigator attempts to make predictions for individuals based on "similar" persons treated with alternative therapies. However, it is rarely the case that everyone in a clinical trial is influenced by the treatment in the same way. An attempt to reduce this heterogeneity of treatment effects (HTE) by using conventional subgroup analysis with one variable at a time is rejected by Kent et al. (2018). They argue that this approach does not work. First, there are many variables on which participants can differ, and there is no way to produce the number of groups that represent these differences. For example, matching on just 20 binary variables would produce over a million groups. Moreover, one would have to start with an enormous sample to maintain adequate statistical power. The authors describe several technical reasons for not recommending this approach to dealing with the HTE problem. They also suggested two other statistical approaches, risk modeling and treatment effect modeling, that may be useful, but more research on both is needed to support their use. Kent et al. (2018) briefly discussed using observational or non-RCT data, but they pointed out the typical problems of missing data and other data quality issues as well as the difficulty in making causal attributions. Moreover, they reiterated their support for the RCT as the "gold standard." Although published in 2018, their article mentioned machine learning only as a question for future research—a question that I address later in this paper. I will also present other statistical approaches to solving the limitations of RCTs.

There is another problem in depending upon RCTs as the gold standard. Nadin (2017) pointed out that failed reproducibility occurs almost exclusively in life sciences, in contrast to the physical sciences. I would add that the behavioral sciences have not been immune from criticisms about replicability. The Open Science Collaboration (2015) systematically sampled 100 results from three top-tier journals in psychology, and only 36% of the replication efforts yielded significant findings. This issue is far from resolved, and it is much more complex than simple replication (Laraway et al. 2019). Nadin (2017) considered the issue of the replicability as evidence of an underlying false assumption about treating humans as if they were mechanistic physical objects and not reactive human beings. He noted that physics is nomothetic, while biology is idiographic, meaning that the former is the study of the formulation of universal laws and the latter deals with the study of individual cases or events.

The Problem of Lack of Learning Through Feedback

Without accurate feedback, there is little learning (Kluger and DeNisi 1996). Clinicians are in a low feedback occupation, and unlike carpenters or surgeons, they are unlikely to get direct accurate feedback on the effects of their activities. When carpenters cut something too short, they can quickly see that it no longer fits and have to start with a new piece, so they typically follow the maxim of measure twice, cut once. Because clinicians in the real world of treatment do not get direct accurate feedback on client outcomes, especially after clients leave treatment, then they are unlikely to learn how to become more effective clinicians from practice alone. Clinical practice is thus similar to an archer's trying to improve while practicing blindfolded (Bickman 1999). Moreover, the services research field does not learn from treatment as usual in the real world, where most treatment occurs, because very few services collect outcome data, let alone try to tie these data to clinician actions (Bickman 2008b).

There are two critical requirements needed for learning. The first is the collection of fine-grained data that are contemporaneous with treatment. The second is the feedback of these data to the clinician or others so that they can learn from these data. Learning can be accomplished with routine use of measures such as patient outcome measures (POMs) and feedback through progress monitoring, measurement-based care (MBC), and measurement feedback systems (MFS). These measurement feedback concepts have repeatedly demonstrated their ability to improve outcomes in therapy across treatment type and patient populations (Brattland et al. 2018; Bickman et al. 2011; Dyer et al. 2016; Gibbons 2015; Gondek et al. 2016; Lambert et al. 2018). Despite this evidence base, most clinicians do not use these measurement feedback systems. For example, in one of the largest surveys of Canadian psychologists, only 12% were using a progress monitoring measure (Ionita et al. 2016).

A Canadian Psychological Association task force (Tasca et al. 2019) reinforced the need for psychologists to systematically monitor and evaluate their services using continuous monitoring and feedback. They stated that the association should encourage regulatory bodies to prioritize training in their continuing education and quality assurance requirements. Moreover, Lewis et al., in their review of measurement-based care (2019), presented a 10-point research agenda that captures much the ideas in the present paper:

> (1) harmonize terminology and specify MBC's core components; (2) develop criterion standard methods for monitoring fidelity and reporting quality of implementation; (3) develop algorithms for MBC to guide psychotherapy; (4) test putative mechanisms of change, particularly for psychotherapy; (5) develop brief and psychometrically strong measures for use in combination; (6) assess the critical timing of administration needed to optimize patient outcomes; (7) streamline measurement feedback systems to include only key ingredients and enhance electronic health record interoperability; (8) identify discrete strategies to support implementation; (9) make evidence-based policy decisions; and (10) align reimbursement structures. (p. 324)

It is not surprising that the measurement feedback approach has not yet produced dramatic effects, given how little we know about what data to collect, how often it should be collected, what feedback should be, and when and how it should be provided (Bickman et al. 2015). Regardless, every time a client is treated, it is an opportunity to learn how to be more effective. By not collecting and analyzing information from usual care settings, we are missing a major opportunity to learn from ordinary services. The most successful model I know of using this real-world services approach is the treatment of childhood cancers in hospitals where most children enter a treatment RCT (O'Leary et al. 2008). These authors note that in the past 50 years, the survival rates for childhood cancer have climbed from 10% to almost 80%. They attribute this remarkable improvement to clinical research through pediatric cooperative groups. This level of cooperation is not easy to develop, and it is not frequently found in mental health services.

The Problem of Insufficiency of Treatment Precision

Most previous research shows differential outcomes among different types of therapies that are minor at most (Wampold and Imel 2015). For example, Weisz et al. (2017) report that in their meta-analysis, the effect of treatment type as a moderator was not statistically significant but there was a significant, but not clearly understood, treatment type by informant interaction effect. In addition, the evidence that therapists have a major influence on the outcomes of psychotherapy is still being hotly debated. The fact that the efficacy of therapists is far from a settled issue is troubling (Anderson et al. 2016; Goodyear et al. 2017; Hill et al. 2017; King and Bickman 2017). Also, current drug treatment choices in psychiatry are successful in only about 50% of the patients (Bzdok and Meyer-Lindenberg 2018) and are as low as 11–30% for antidepressants (Dwyer et al. 2018). While antidepressants are more effective than placebos, they have small effect sizes (Perlis 2016), and the choice of specific medicine is a matter of trial and error in many cases.

It is relatively easy to distinguish one type of drug from another but not so for services, where even dosage in psychosocial treatments is hard to define. According to Dwyer et al. (2018), "Currently, there are no objective, personalized methods to choose among multiple options when tailoring optimal psychotherapeutic and pharmacological treatment" (p. 105). A recent summary concluded that after 46 years and 57 studies, it is unknown which patients benefit from interpersonal psychotherapy (IPT) versus another treatment (Bernecker et al. 2017). However, to provide a more definitive answer to the question about which treatments are more effective, we need head-to-head direct comparisons between different treatments and network meta-analytic approaches such as those used by Dagnea et al. (2016). The field of mental health is not alone in finding that many popular medications do not work with most of the people who take them. Nexium, a common drug for treating heartburn, works with only 1 person out of 25, while Crestor, used to treat high cholesterol, works with only 1 out of 20 (Schork 2015). This poor alignment between what the patient needs, and the treatment provided is the primary basis for calling for a more precise medicine approach. This lack of precision leads to the application of treatments to people who cannot benefit from it, thus leading to overall poor effectiveness.

In summary, a deep and growing body of work has led me to conclude that we need additional viable approaches to a RCT when it comes to conducting services-related research. An absence of rigorous evaluation of treatments that are usually provided in the community contributes to a gap in our understanding why treatments are ineffective (Bickman 2008b). Poor use of measurement in routine care (Bickman 2012) and the absence of measurement feedback systems and clinician training and supervision (Garland et al. 2010) are rampant. There also a dire need for the application of more advanced analytics and data mining techniques in the mental health services area (Bickman et al. 2016). These and other such challenges have in turn informed my current thinking about alternative or ancillary approaches for addressing the multitude of problems plaguing the field of mental health services.

An Introduction to Artificial Intelligence (AI) and Precision Medicine

The five problems I have described above constitute significant obstacles to achieving accessibility, efficiency, and effectiveness in mental health services. Nevertheless, there is a path forward that I believe can help us reach these goals. Artificial intelligence promises to transform the way healthcare is delivered. The core of my recommendations in this paper rests on the revolutionary possibilities of artificial intelligence for improving mental healthcare through precision medicine that allows us to take into account the individual variability that exists with respect to genetic and other biological, environmental, and lifestyle characteristics. Several others have similarly signaled a need for considering the use of personalized approaches to service delivery. For example, Weisz and his colleagues (Marchette and Weisz 2017; Ng and Weisz 2016) called for more idiographic research and for studies tailoring strategies in usual care. Kazdin (2019) focused on expanding mental health services through novel models of intervention delivery; called for task shifting among providers; advocated designing and implementing treatments that are more feasible, using disruptive technologies, for example, smartphones, social media such as Twitter and Facebook, and socially assistive robots; and emphasized social network interventions to connect with similar people.

AI is currently used in areas ranging from prediction of weather patterns to manufacturing, logistic planning to determine efficient delivery routes, banking, and stock trading. AI is used in smartphones, cars, planes, and the digital assistants Siri and Alexa. In healthcare, decision support, testing and diagnosis, and self-care also use AI. AI can sort through large data sets and uncover relationships that humans cannot perceive. Through learning that occurs with repeated, rapid use, AI surpasses the abilities of humans only in some areas. However, I would caution potential users that there are significant limitations associated with AI that are discussed later in this paper. Rudin and Carlson (2019) present a non-technical and well written review of how to utilize AI and of some of the problems that are typically encountered.

Varieties of AI: A Basic Introduction

AI is not one type of program or algorithm. Machine learning (ML), a major type of AI, is the construction of algorithms that can learn from and make predictions based on data. It can be (1) supervised, in which the outcome is known and labeled by humans and the algorithm learns to get that outcome; (2) unsupervised, when the program learns from data to predict specific outcomes likely to come from the patterns identified; and (3) reinforcement learning, in which ML is trial and error. In most cases, there is an extensive training data set that the algorithm "learns" from, followed by an independent validation sample that tests the validity of the algorithm. Other variations of AI include random forest, decision trees, and the support vector machine (SVM), a multivariate supervised learning technique that classifies individuals into groups (Dwyer et al. 2018; Shrivastava et al. 2019). The latter is most widely used in psychology and psychiatry. Artificial neural networks (ANNs) or "neural networks" (NNs) are learning algorithms that are conceptuality related to biological neural networks. This approach can have many hidden layers. Deep learning is a special type of machine learning. It helps to build learning algorithms that can function conceptually in a way similar to the functioning of the human brain. Large amounts of data are required to use deep learning. IBM's Watson won *Jeopardy* with DeepQA algorithms designed for question answering. As exemplified by the term *neural networks*, algorithm developers appear to name their different approaches with reference to some biological process. Genetic algorithms are based on the biological process of gene propagation and the methods of natural selection, and they try to mimic the process of natural evolution at the genotype level. It has been a widely used approach since the 1960s.

Natural language processing (NLP) involves speech recognition, natural language understanding, and natural language generation. NLP may be especially useful in analyzing recordings of a therapy session or a therapist's notes. Affective computing or sentiment analysis involves the emotion recognition, modeling, and expression of emotion by robots or chatbots. Sentiment analysis can recognize and respond to human emotions. Virtual reality and augmented reality are human–computer interfaces that allow a user to become immersed within and interact with computer-generated simulated environments.

Hinton (2018), a major contributor to research on AI and health, described AI as the use of algorithms and software to approximate human cognition in the analysis of complex data without being explicitly programmed. The primary aim of health-related AI applications is to analyze relationships between prevention or treatment techniques and patient outcomes. AI programs have been developed and applied to practices such as diagnosis processes, treatment protocol development, drug development, personalized medicine, and patient monitoring and care. Deep learning is best at modeling very complicated relationships between input and outputs and all their interactions, and it sometimes requires a very large number of cases—in the thousands or tens of thousands—to learn. However, there appears to be no consensus about how to determine, a priori, the number of cases needed, because the number is highly dependent on the nature of the problem and the characteristics of the data.

Uses of AI in Medicine

AI is already widely used in medicine. For example, in ophthalmology, photos of the eyes of persons with diabetes were screened with 94% specificity and 98% sensitivity in detecting diabetes (Gargeya and Leng 2017). One of the more prolific uses of AI is in the diagnosis of skin cancer. In a study that scanned 129,450 clinical images, the AI approach had accuracy similar to that of board-certified dermatologists (Esteva et al. 2017). Cardiovascular risk prediction with ML is significantly improved over established methods of risk prediction (Krittanawong et al. 2019; Weng et al. 2017). However, a study by Desai et al. (2020) found only limited improvements in predicting heart failure over traditional logistic regression. In cancer diagnostics, AI identified malignant tumors with 89% accuracy compared to 73% accuracy for human pathologists (Liu et al. 2017). The IBM's Watson AI platform took only 10 min to analyze a genome of a patient with brain cancer and suggest a treatment plan, while human experts took 160 h (Wrzeszczynski et al. 2017).

AI has also been used to develop personalized immunotherapy for cancer treatment (Kiyotani et al. 2018). Rajpurkar et al. (2017) compared 50 chest X-rays for signs of pneumonia using a state-of-the-art 121-layer convolutional neural network (CNN) program with a "swarm" of radiologists (groups connected by swarm algorithms) and found the latter to be significantly more accurate. In a direct comparison between 101 radiologists on 28,296 interpretations and a stand-alone deep learning AI program designed to detect breast cancer in mammography, the AI program was as accurate as the radiologists (Rodriguez-Ruiz et al. 2019).

As Topol (2019b) noted, AI is not always the winner in comparison with human experts. Moreover, many of these applications have not been used in the real world, so we do not know how well AI will scale up in practice. Topol describes other concerns with AI, many of which are discussed later in this paper. Finally, many of the applications are visual, such as pictures of skin or scans, for which AI is particularly well suited. Large banks of pictures often form the training and testing data for this approach. In mental health, visual data are not currently as relevant. However, there is starting to be some research on facial expressions in diagnosing mental illness. For example, Abdullah and Choudhury (2018) cite several studies that showed that patients with schizophrenia tend to show reduced facial expressivity or that facial features can be used to indicate mental health status. More generally, there is research showing how facial expressions can be used to indicate stress

(Mayo and Heilig 2019). Visual data are ripe for exploration using AI.

Although an exhaustive review of the AI literature and its applications is well beyond the focus of this paper, Rudin and Carlson (2019) present a well-written, non-technical review of how to utilize AI and of some of the problems that are typically encountered. Topol (2019a), in his book titled *Deep Medicine: How Artificial Intelligence Can Make Healthcare Human Again,* includes a chapter on how to use of AI in mental health. Topol (2019b) also provides an excellent review of AI and its application to health and mental health in a briefer format. Buskirk et al. (2018) and Y. Liu et al. (2019) provide well-written and relatively brief introductions to ML's basic concepts and methods and how they are evaluated. A more detailed introduction to deep learning and neural networks is provided by Minar and Naher (2018). In most cases, I will use the generic term *AI* to refer to all types of AI unless the specific type of AI (e.g., ML for machine learning, DL for deep learning, and DNN for deep neural networks) is specified.

Precision Medicine

Precision medicine has been defined as the customization of healthcare, with medical decisions, treatments, practices, or products being tailored to the individual patient (Love-Koh et al. 2018). Typically, diagnostic testing is used for selecting the appropriate and best therapies based a person's genetic makeup or other analysis. In an idealized scenario, a person may be monitored with hundreds of inputs from various sources that use AI to make predictions. The hope is that precision medicine will replace annual doctor visits and their granular risk factors with individualized profiles and continuous longitudinal health monitoring (Gambhir et al. 2018). The aim of precision medicine, as stated by President Barack Obama when announcing his precision medicine initiative, is to find the long-sought goal of "delivering the right treatments, at the right time, every time to the right person" (Kaiser 2015).

Both AI and precision medicine can be considered revolutionary in the delivery of healthcare, since they enable us to move from one-size-fits-all diagnoses and treatment to individualized diagnoses and treatments that are based on vast amounts of data collected in healthcare settings. The use of AI and precision medicine to guide clinicians will change diagnoses and treatments in significant ways that will go beyond our dependence on the traditional RCT. Precision medicine should also be seen as evolutionary since even Hippocrates advocated personalizing medicine (Kohler 2018).

The importance of a precision medicine approach was recognized in the field of prevention science with a special issue of *Prevention Science* devoted to that topic (August and Gewirtz 2019). The articles in this special issue recognize the importance of identifying moderators of treatment that predict heterogeneous responses to treatment. Describing moderators is a key feature of precision medicine. Once these variables are discovered, it becomes possible to develop decision support systems that assist the provider (or even do the treatment assignment) in selecting the most appropriate treatment for each individual. This general approach has been tried using a sequential multiple assignment randomized trial (SMART) in which participants are randomized two to three times at key decision points (August et al. 2016). What I find notable about this special issue is the absence of any focus on AI. The articles were based on a conference in October 2016, and apparently the relevance of AI had not yet influenced these very creative and thoughtful researchers at that point.

Precision medicine does not have an easy path to follow. X. Liu et al. (2019b) describe several challenges, including the following three. Large parts of the human genome are not well enough known to support analyses; for example, almost 90% of our genetic code is unknown. It is also clear that a successful precision medicine approach depends on having access to large amounts of data at multiple levels, from the genetic to the behavioral. Moreover, these data would have be placed into libraries that allow access for researchers. The U.S. federal government has a goal of establishing such a library with data on one million people through NIH's All of Us Research Program (https://allofus.nih.gov/). Recruitment of volunteers who would be willing to provide data and the "harmonization" of data from many different sources are major issues. X. Liu et al. (2019b) also point to ethical issues that confront precision medicine, such as informed consent, privacy, and predictions that someone may develop a disease. These issues are discussed later in this paper.

Chanfreau-Coffinier et al. (2019) provided a helpful illustration of how precision medicine could be implemented. They convened a conference of 80 Veterans Affairs stakeholders to develop a detailed logic model that can be used by an organization planning to introduce precision medicine. This model includes components typically found in logic models, such as inputs (clinical and information technology), big data (analytics, data sources), resources (workforce, funding) activities (research), outcomes (healthcare utilization), and impacts (access). The paper also includes challenges to implementing precision medicine (e.g., a poorly trained workforce) that apply to mental health.

AI has the potential to unscramble traditional and new diagnostic categories based on analysis of biological/genetic and psychological data, and in addition, more data will likely be generated now that the potential for analysis has become so much greater. AI also has the potential to pinpoint those individuals who have the highest probability of benefiting from specific treatments and to provide early indicators of

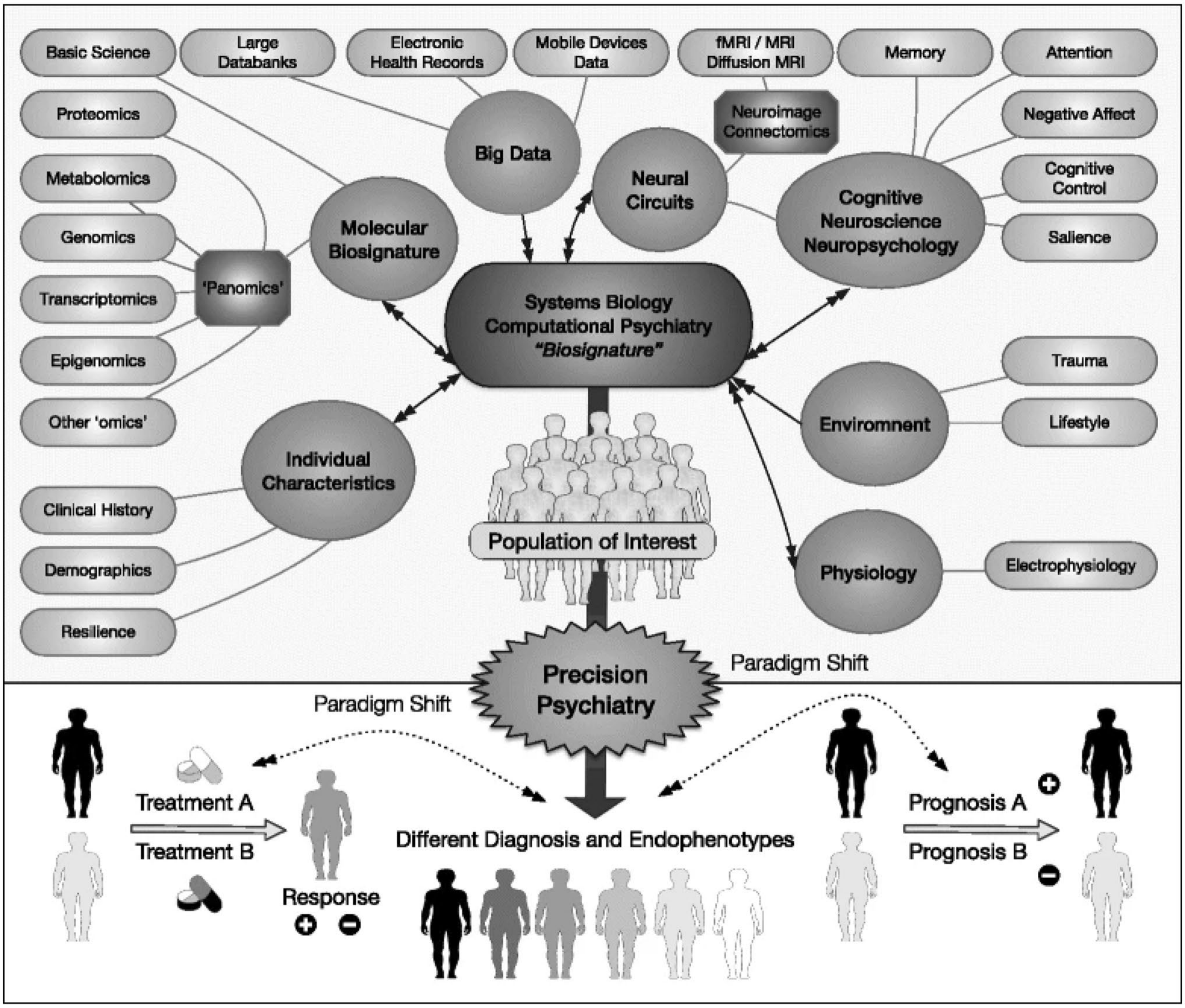

Fig. 1 Domains related to precision psychiatry. Source: reprinted from Fernandes et al. (2017). Distributed under Creative Commons Attribution 4.0 International License

success or failure of treatment. Research is currently being undertaken to provide feedback to clinicians at key decision points as an early warning of relapse.

Precision Psychiatry

Fernandes et al. (2017) describe what the authors call the *domains* related to precision psychiatry (see Fig. 1). These domains include many approaches and techniques, such as panomics, neuroimaging, cognition, and clinical characteristics, that form several domains including big data and molecular biosignature; the latter includes biomarkers. The authors include data from electronic health records, but I would also include data collected from treatment or therapy sessions as well as data collected outside of these sessions. These domains can be analyzed using biological and computational tools to produce a biosignature, a higher order domain that includes data from all the lower level techniques and approaches. This set of biomarkers in the biosignature should result in improved diagnosis, classification, and prognosis, as well as individualized interventions. The authors note that this bottom-up approach, from specific approaches to domains to the ultimate biosignature, can also be revised to a top-down approach, with the biosignature studied to better understand domains and its specific components. The bottom of the figure shows a paradigm shift where precision psychiatry contributes to different treatments being applied to persons with different diagnoses and *endophenotypes*, producing different prognoses. *Endophenotypes* is a term used in genetic epidemiology to separate different behavioral

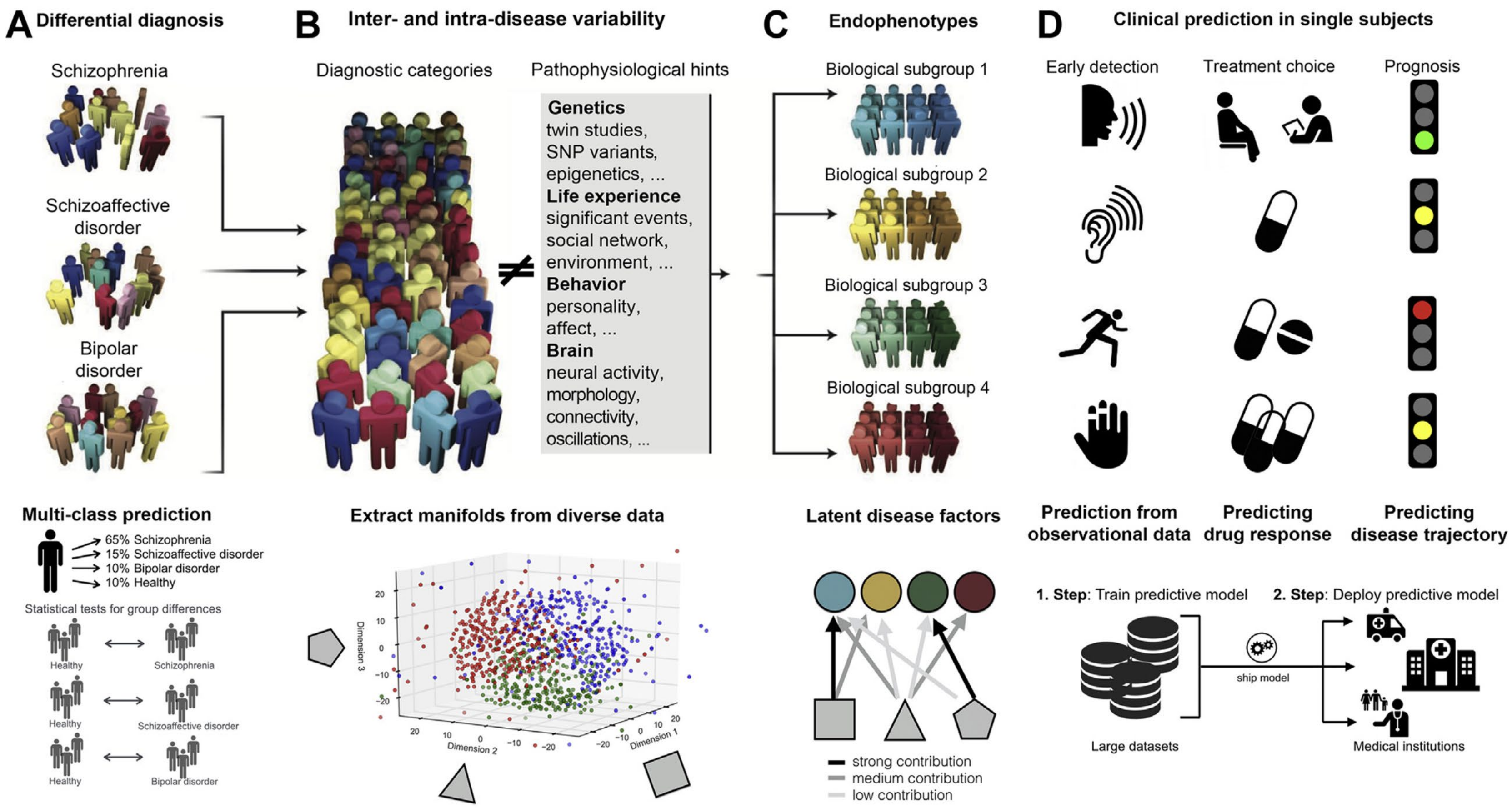

Fig. 2 Model of precision psychiatry. Source: reprinted from Bzdok and Meyer-Lindenberg (2018). Used by permission of Elsevier: http://www.elsevier.com

symptoms into stable phenotypes with a well-defined genetic relationship (Fig. 2).

Another perspective on precision psychiatry is presented by Bzdock and Meyer-Lindberg (2018). Both models contain similar concepts. Both start with a group of persons containing multiple traditional diagnoses. Bzdock and Meyer-Lindberg recognize that these psychiatric diagnoses are often artificial dichotomies. Machine learning is applied to diverse data from many sources and extracts hidden relationships. This produces different subgroups of endophenotypes. Machine learning is also used to produce predictive models of the effects of different treatments instead of the more typical trial and error. Further refinement of the predictive ML models results in better treatment selection and better prediction of the disease trajectory. An excellent overview of deep neural networks (DNNs) in psychiatry and its applications is provided by Durstewitz et al. (2019). In addition to explaining how DNNs work, they provide some suggestions on how DNNs can be used in clinical practice with smartphones and large data sets. A major feature of deep neural networks is their ability to learn and adapt with experience. While DNNs typically outperform ML, the authors state that they do not fully understand why this is the case. In mental health, DNNs have been mostly used in diagnosis and predictions but not in designing personalized treatments. DNN's ability to integrate many different data sets (e.g., various neuroimaging data, movement patterns, social media, and genomics) should provide important insights on how to personalize treatments. Regardless of the model used, Eyre et al. (2020) remind us that consumers should not be left out of the development of precision psychiatry.

Precision Mental Health Services

In my conceptualization of precision medicine, precision mental health encompasses precision psychiatry and any other precision approach such as social work that focuses on mental health (Bickman et al. 2016). There has not been much written about using a precision approach with psychosocial mental health services. Possibly it is psychiatry's close relationship to general medicine and its roots in biology that make psychiatry more amenable to the precision science approach. In addition, the use of the precision construct is being applied in other fields, as exemplified by the special issue of the *Journal of School Psychology* devoted to precision education (Cook et al. 2018) and precision public health (Kee and Taylor-Robinson 2020). However, in this paper I am primarily addressing the use of psychosocial treatment of mental health problems, which differs in important ways from psychiatric treatment. For example, precision psychosocial mental health treatment does not have a strong biological/medical perspective and does not focus almost exclusively on medication; instead, it emphasizes psychosocial interventions. Psychosocial mental health services are also provided in hospital settings, but their primary use is in community-based services. These differences lead to different data sources for AI analyses. It is highly unlikely that electronic mental healthcare records found outside of hospital settings contain biological and genomic data (Serretti 2018). But hospital records are not likely to contain the detailed treatment process data that could possibly be found in community settings. The genomic and biological data offer new perspectives but may not be informative until we have a better understanding about the genomic basis of mental illness. In addition, the internet of things and smart healthcare connect wearable and home-based sensors that can be used to monitor movement, heart rate, ECG, EMG, oxygen level, sleep, and blood glucose, through wi-fi, Bluetooth, and related technologies. (Sundaravadivel et al. 2018). With wider use of very fast 5G internet service, there will be a major increase in the growth of the internet of things.

I want to emphasize that applying precision medicine concepts to mental health services, especially psychotherapy, is a very difficult undertaking. The data requirements for psychosocial mental health treatment are more similar to meteorology or weather forecasting than to agriculture, which is considered the origin of the RCT design. People's affect, cognition, and behavior are constantly changing just like the variables that affect weather. But unlike meteorology, which is mainly descriptive and not yet engaged in interventions, mental health services are interventions. Thus, in addition to client data, we must identify the variables that are critical to the success of the intervention. We are beginning to grasp how difficult this task is as we develop greater understanding that the mere labeling of different forms of treatment by location (e.g., hospital or outpatient) or by generic type (e.g., cognitive behavior therapy) is not sufficiently informative. Moreover, the emergence of implementation sciences has forced us to face the fact that a treatment manual describes only some aspects of the treatments as intended but does not describe the treatment that is actually delivered. NLP is a step in the right direction in trying to capture some aspects of treatment as actually delivered.

Data quality is the foundation upon which AI systems are built. While medical records are of higher technical quality than community-based data because they must adhere to national standards, I believe that the nascent interest in measurement-based care and measurement feedback systems in community settings bodes well for improved data systems in the future. Moreover, although electronic hospital-based data may be high quality from a technical viewpoint (validity, reliability) and be very large, they probably do not contain the data that are valuable for developing and evaluating mental health services. The development of electronic computer-based data collection and feedback systems will become more common as the growth in AI demands large amounts of good-quality treatment and finer grained longitudinal outcome data. There is a potential reciprocal relationship between the AI needs for large, high-quality data sets and the development of new measurement approaches and the electronic systems needed to collect such data (Bickman 2008a; Bickman et al. 2012a, 2016). To accomplish this with sufficiently unbiased and valid data will be a challenge.

Solutions to the Five Problems

Solutions to the Problem of the Diagnosis Muddle

AI can bypass many definitional problems by not using established diagnostic systems. ML can use a range of variables to describe the individual ML classifier systems (Tandon and Tandon 2018). Moreover, additional sources of data that help in classification are now feasible. For example, automated analysis of social media including tweets and Facebook can detect depression, with accuracy measured by area under the curve (AUC) ranging from 0.62 to 0.74 compared to clinical interviews with AUCs of 0.90 (Guntuku et al. 2017). As noted earlier, DNNs have been shown to be superior to other machine learning approaches in general and specifically in identifying psychiatric stressors for suicide from social media (Du et al. 2018). Predictions of 1,479 adolescent suicides with ML showed high accuracy (AUC > 0.80) and outperformed traditional logistic regression analyses (0.5–0.6 AUCs) (Tandon and Tandon 2018). Saxe has published a pioneering proof of concept that has demonstrated that ML methods can be used to predict child posttraumatic stress (Saxe et al. 2017). ML was more accurate than humans in predicting social and occupational disability with persons in high-risk states of psychosis or with recent-onset depression (Koutsouleris et al. 2018a). Machine learning has also been used in predicting psychosis using everyday language (Rezaii et al. 2019).

Another application of AI to diagnosis is provided by Kasthurirathne et al. (2019). They demonstrated the ability to automate screening for 84,317 adult patients in need of advanced care for depression using structured and unstructured data sets covering acute and chronic conditions, patient demographics, behaviors, and past service use history. The use of many existing data elements is a key feature and thus does not depend on single screening instruments. The authors used this information to accurately predict the need for advanced care for depression using random forest classification ML.

Milne et al. (2019) recognized that in implementing online peer counseling, professionals need to participate and/or provide safety monitoring in using AI. However, cost and scalability issues appeared to be insurmountable barriers. What is needed is an automated triage system that would direct human moderators to cases that require the most urgent attention. The triage system Milne et al. developed sent human moderators color-coded messages about their need to intervene. The algorithm supporting this triage system was based on supervised ML. The accuracy of the system was evaluated by comparing a test set of manually prioritized messages with the ones developed through the algorithm. They used several methods to judge accuracy, but their main one was an f-measure**,** or the harmonic mean of recall (i.e., sensitivity) and precision (i.e., positive predictive value). Regression analysis indicated that the triage system made a significant and unique contribution to reducing the time taken to respond to some messages, after accounting for moderator and community activity. I can see the potential for this and similar AI approaches to deal with the typical service setting where some degree of supervision is required but even intermittent supervision is not feasible or possible.

Another use of ML as a classification tool is provided by Pigoni et al. (2019). In their review of treatment resistant depression, they found that ML could be used successfully to classify responders from non-responders. This suggested that stratification of patients might help in selecting the appropriate treatment, thus avoiding giving patients treatments that are unlikely to work with them. A more general systematic review and meta-analysis of the use of ML to predict depression are provided by Lee et al. (2018). The authors found 26 qualitative and 20 quantitative studies that qualified for inclusion in their review. While most of the studies were retrospective, they did find predictions with an average overall accuracy of 0.82.

Kaur and Sharma (2019) reviewed the literature on diagnosis of ten different psychological disorders and examined the 16 different data mining and software approaches (AI) used in 126 different publications. Depending on the disorder and the software used, the accuracy ranged from 84 to 98%. Accuracy was defined differently depending on the study. Only 1% of the articles exploring diagnosis of any health problem were found to be for psychological problems. This suggests that we need more studies on diagnosis and AI. A very informative synthesis and review are provided by Low et al. (2020). They screened 1395 studies and reviewed the 127 that met the inclusion criterion: studies from the last 10 years using speech to identify the presence or severity of disorders through ML methods. They concluded that ML could be predictive, but confidence in any conclusions was dampened by the general lack of cross-validation procedures. The article contains very useful information on how best to collect and analyze speech samples.

Another innovative approach using ML focused on wearable motion detector sensors, in which these devices were worn for 20s during a 90-s mood induction task (seeing a fake snake). These data were able to distinguish children with an internalizing disorder from controls with 81% accuracy (McGinnis et al. 2019). This approach has potential for screening children for this disorder.

A problem that seemingly has been ignored by most studies that deal with classification or diagnosis is the gold standard by which accuracy is judged. In most cases, the gold standard is human judgment, which is especially fallible when it comes to mental health diagnosis. We can clearly measure whether the AI approach is faster and less expensive than human judgment, but is the ultimate in AI accuracy matching human judgment with all its flaws? I believe that the endpoint that must also be measured is client clinical mental health improvement. A system that provides faster and less expensive diagnosis but does not lead to more precise treatment and better clinical outcomes will save us time and money, which are important, but they will not be the breakthrough for which we are looking.

A solution to the problems described above will involve the integration of causal discovery methods with AI approaches. AI methods are capable of improving our capacity to predict outcomes. To enhance predictability, we will need to identify the factors in the predictive models that are causal. Thus, there is the need to identify techniques that provide us with causal knowledge, which currently is based primarily on RCTs. But, for real-world and ethical reasons, human etiological experiments can rarely be conducted. Fortunately, there are newer AI methods that can be used to infer causes, which include well validated tests of conditional independencies based on the Causal Markov Condition (Pearl 2009; Aliferis et al. 2010; Saxe 2019).

These methods have been successfully used outside of psychiatry (Sachs et al. 2005; Ramsey et al. 2016; Statnikov et al. 2005) and have, in the last five years, been applied in research on mental health, largely by the team of Glenn Saxe at New York University and Constantin Aliferis and Sisi Ma at University of Minnesota. This group has reported causal models of PTSD in hospitalized injured children (Saxe et al. 2016, 2017), children seen in outpatient trauma centers (Saxe et al. 2016), maltreated children (Morales et al. 2018), adults seen in emergency rooms (Galatzer-Levy et al. 2017), and police officers who were exposed to trauma (Saxe et al. in press). Saxe (2020) recently described the promise of these methods for psychiatric diagnosis and personalized precision medicine.

Solutions to the Problem of Poorly Designed Measures

New measures need to be developed that cover multiple domains of mental health, are reported by different respondents (e.g., child, parent, clinician), and are very brief. Cohen (2019) provides an excellent overview of what he calls *ambulatory biobehavioral technologies* in a special section of *Psychological Assessment*. He notes that the development of mobile devices can have a major impact on psychological assessment. He cautions, however, that while some of these approaches have been used for decades, they still have not progressed beyond the proof of concept phase for clinical and commercial applications.

Ecological Momentary Assessments

Ecological momentary assessment (EMA) is a relatively new approach to measurement development. EMA is the collection of real-time data collected in naturalistic environments. This approach uses a wide range of smart watches, bands, garments, and patches with embedded sensors (Gharani et al. 2017; Pistorius 2017). For example, using smartphones, researchers have identified gait features for estimating blood alcohol content level (Gharani et al. 2017). Other researchers have been able to map changes in emotional state ranging from sad to happy by using a movement sensor on smart watches (Quiroz et al. 2018). Others have described real-time fluctuations in suicidal ideation and its risk factors, using an average of 2.5 assessments per day (Kleiman et al. 2017). Social anxiety has been assessed from global positioning data obtained from smart watches by noting that socially anxious students were found to avoid public places and to spend more time at home than in leisure activities outside the home (Boukhechba et al. 2018). A review of 42 studies using EMA concluded that the compliance rate was moderate but not optimal and could be affected by study design (Wen et al. 2017). This review is also a good source of descriptions of different approaches to using EMA. Another good summary that focused on EMA in the treatment of psychotic disorders can be found in Bell et al. (2017). For EMA use in depression and anxiety, Schueller et al. (2017) is a good source.

EMA has been used to measure cardiorespiratory function, movement patterns, sweat analysis, tissue oxygenation, sleep, and emotional state (Peake et al. 2018). Harari et al. (2017) present a catalog of behavior in more than 50 aspects of daily living that can be used in studying physical movement, social interactions, and daily activities. These include walking, speaking, text messaging, and so on. These all can be collected from smartphones and serve as an alternative to traditional survey approaches. However, it is still not clear what higher-level constructs are measured using these approaches.

A comprehensive and in-depth review of 127 studies that have used speech to assess psychiatric disorders is provided by Low et al. (2020). They conclude that speech processing technology could assist in mental health assessments but believe that there are many obstacles to this use, including the need for longitudinal studies. Another interesting application for children is the use of inexpensive screening for internalizing disorders. McGinnis et al. (2019) monitored the child's motion for 20s using a commercially available and inexpensive wearable sensor. Using a supervised ML approach, they obtained an 81% accuracy (67% sensitivity, 88% specificity) compared to similar clinical threshold on parent-reported child symptoms that differentiate children with an internalizing diagnosis from controls without such a diagnosis.

In a systematic review of EMA use in major depression, Colombo et al. (2019) evaluated 33 studies that met their criteria for inclusion. These studies measured a wide variety of variables including self-reported symptoms, sleep patterns, social contacts, cortisol, heart rate, and affect. They point out many of the advantages of using EMAs such as real-time assessments, capturing the dynamic nature of change, improving generalizability, and providing information about context. They believe that the use of EMAs has resulted in novel insights about the nature of depression. They do note that there are few evaluations of these measures, and there is not much use in actual clinical practice.

Mohr et al. (2017) note that most of the research on EMA has been carried out primarily by computer scientists and engineers using a very different research model than social and behavioral scientists. While computer scientists are mostly interested in exploratory proof of concepts approach (does it work at all?) using very small samples, social/behavioral scientists are more typically theory driven and investigate under what conditions the intervention will work.

Measuring Content of Treatment

Mental health care, apart from medication, is almost exclusively verbal. Several approaches have been tried to capture the content of treatment sessions. My colleagues and I have tried by asking clinicians to use a brief checklist of topics discussed after each therapy session (Kelley et al. 2012). Although this technique produced some interesting findings such as the identification of topics that the clinician did not discuss but that were believed to be important by the youth or parent, it is clearly filtered by what the clinician recalls and is willing to check off as having been discussed. While recordings provide a richer source of information, coding recordings manually is too expensive and slow for the real world of service delivery. The content of therapy sessions, including notes kept by clinicians, is pretty much ignored by researchers because of the difficulty and cost of manually coding those sources. However, advances in natural language processing (NLP) are now being explored as a way of capturing aspects of the content of therapy sessions. For example, Tanana et al. (2016) have shown how two types of NLP techniques can be used to study and code the use of motivational interviewing in taped sessions. Carcone et al. (2019) also showed that they could accurately code motivational interviewing (MI) clinical encounter transcripts with sufficient accuracy. Other researchers have used AI to analyze speech to distinguish between what they called high- and low-quality counselors (Pérez-Rosas et al. 2019). Some colleagues and I have submitted a proposal to NIMH to refine NLP tools that can be used to supervise clinicians implementing an evidence-based treatment using AI. As far as we know, using NLP to measure fidelity and provide feedback to clinicians has not been studied in a systematic way.

While AI appears to be an attractive approach to new ways of analyzing data, it should be noted that, as always, the quality of the analysis is highly dependent on the quality of the data. Jacobucci and Grimm (2020) caution us that "In psychology specifically, the impact of machine learning has not been commensurate with what one would expect given the complexity of algorithms and their ability to capture nonlinear and interactive effects" (p. 1). One observation made by these authors is that the apparent lack of progress in using AI may be caused by "throwing the same set of poorly measured variables that have been analyzed previously into machine learning algorithms" (p. 2). They note that this is more than the generic garbage in, garbage out problem, but it is specifically related to measurement error, which can be measured relatively accurately.

Solutions to the Problem of the Primacy of RCTs

As described earlier, our privileging of RCTs has contributed to a lack of focus on a precision approach to mental health services. This has resulted in the problem of ignoring the clinical need for predicting for an individual in contrast to establishing group difference, the approach favored by the experimentalist/ hypothesis testing tradition. AI offers an approach to the discovery of important relationships in mental health in addition to RCTs that are based on single-subject prediction accuracy and not null hypothesis testing (Bzdok and Karrer 2018). Saxe et al. (2016) have demonstrated the use of the Complex-Systems-Causal Network method to detect causal relationships among 111 variables and 167 bivariate relations in a psychiatric study using algorithms. A comprehensive review and meta-analysis of machine learning algorithms that predict outcomes of depression showed excellent accuracy (0.82) using multiple forms of data (Lee et al. 2018). It is interesting to note that none of the 26 scholars commenting on the RCT special issue in *Social Science and Medicine* (Deaton and Cartwright 2018) specifically mentioned the use of AI as a potential solution to some of the problems of using average treatment effects (ATEs).

Kessler et al. (2019a) noted that clinical trials do not tell us which treatments are more effective for which patients. They suggested that what they label as precision treatment rules (PTRs) be developed that are predictors of the relative treatment effectiveness of different treatments. The authors presented a comprehensive discussion on how to use ML to develop PTRs. They concluded that the sample sizes needed are much larger than usually those found in RCTs; observational data, especially from electronic medical records (EMRs) can be used to deal with the sample size issue; and statistical methods can be used to balance both observed and unobserved covariates using instrumental variables and discontinuity designs. They do note the difficulty in obtaining full baseline data from EMRs and suggest several solutions for this problem, including supplemental data collection and links to other archival sources. They recommend the use of an ensemble ML approach that combines several algorithms. They are clear that their suggestions are exploratory and require verification, but they are more certain that if ML improves patient outcomes, it will be a substantial improvement.

Wu et al. (2020) collaborated with Kessler on a proof of concept of a similar model called individualized treatment rules (ITR). In a model simulation, they used a large sample ($n = 32{,}277$) with an ensemble ML method to identify the advantages of using ML algorithms to estimate the outcomes if a precision medicine approach was taken in prescribing medication for persons with first-onset schizophrenia. They found that the treatment success was estimated to be 51.7% under ITR compared to 44.5% with the medication that was actually used. Wu et al. see this as a first step that needs to be confirmed by pragmatic RCTs. Kessler et al. (2019b) conducted a relatively small randomized study ($n = 148$) in which soldiers seeking treatment were judged to be at risk for suicide. They were randomly assigned to two types of treatment but not on the basis of any a priori PTR. The data from that study were then analyzed using ML to produce PTRs. These data were then modeled in a simulation to see if the PTR would have produced better outcomes. The authors did find that the simulated PTR produced better effects.

Lenze et al. (2020) address the problems of RCTs from a somewhat different perspective than I have presented here and suggest a potential solution that they call *precision clinical trials* (PCTs). The authors propose that the problem with most existing RCTs is that they measure only the fixed baseline characteristics that are not usually sensitive to detecting treatment responders. Moreover, treatment is typically not dynamically adapting to the client during treatment, and measures are not administered with sufficient frequency. Instead, the PCTs would:

(1) first attempt to determine whether short-term responses to the intervention could determine who was a likely candidate for that specific treatment;
(2) initiate the treatment in an adaptive fashion that could vary over time, using stepped care or just-in-time adaptations that are responsive to the client's changing status, and frequently collect data possibly using Multiple Assignment Randomized Trial methods; and
(3) use frequent precision measurement, possibly using ecological momentary assessments described earlier.

Coincidently, they illustrate the application of PCTs using repetitive transcranial magnetic stimulation (rTMS), a form of brain stimulation therapy used to treat depression and anxiety that has been in use since 1985. rTMS will be described later in connection with what I call a third path for services and AI.

It is disappointing that I could not find any examples of published research that used a RCT to test whether an AI approach to an actual, not simulated, delivery of a mental health treatment produces better clinical outcomes than a competitive treatment or even treatment as usual. This is clearly an area requiring further rigorous empirical investigation.

Solutions to the Problem of the Lack of Learning Through Feedback

Imel et al. (2017) provide an excellent overview on how AI and other technologies can be used for monitoring and feedback in psychotherapy in both training and supervision. Imel et al. (2017) used ML to code and provide data to clinicians on metrics used to measure the quality of motivational interviewing (MI). A prior study (Tanana et al. 2016) established that ML was able to code MI quality metrics with accuracy similar to human coders. They conducted a pilot study using standardized patients and 10-min speech segments that was designed to test the feasibility of providing feedback to 21 clinicians on the quality of their MI intervention. The feedback was not in real-time but was provided after the session. They were able to establish that clinicians thought highly of the feedback they received. The authors anticipate that further developments in this technology will lead to its widespread use in supervision and in real-time feedback. It would seem that the next step is evaluating the enhanced AI feedback procedure in a real-world effectiveness study.

Another example of the use of NLP application is the use of a bot that was trained to assess and provide feedback on specific interviewing and counseling skills such as asking open-ended questions and providing feedback (Tanana et al. 2019). After training the bot on 2345 transcripts, 151 non-therapists (using Amazon Mechanical Turk recruits) were randomly assigned to either immediate feedback on a practice session with the bot or just encouragement on the use of those skills. The group provided the feedback were significantly more likely to use reflection even when feedback was removed. The authors consider this to be a proof of concept demonstration because of the many limitations (e.g., use of non-therapists). A plan for using NLP to monitor and provide feedback to clinicians on the implementation of an evidenced program is provided by Berkel et al. (2019). They provide excellent justification for using NLP to accomplish this goal, but unfortunately it is only a design at this point.

Rosenfeld et al. (2019) see AI making major contributions to improving the quality of treatment through efficient continuous monitoring of patients. Until now, monitoring was limited to in-session contacts or manual contacts, an approach that is not practical or efficient. The almost universal availability of smartphones and other internet active devices (internet of things) makes collecting data from clients practical and efficient. These various data sources provide feedback to providers so that they can predict and prevent relapse and compliance with treatment, especially medication. The authors note that there is not a large body of research in this area, but early studies are positive.

One concrete application of AI to providing feedback is described by Ryan and his colleagues (Ryan et al. 2019). Their article only describes how such could be done; unfortunately, it is not an actual study but a suggestion on how to apply AI for feedback to physicians to improve their communications with patients. They note that routine assessment and feedback are not done manually because of the cost and time requirements. However, AI can automate these tasks by evaluating recordings. They suggest using already existing AI approaches that are in use by call centers to categorize and evaluate communication along the following dimensions: speaker ratio that indicates listening, overlapping talk that are interruptions, pauses longer than two seconds, speed, pitch, and tone. The content could also be evaluated along the dimensions of the use of plain language, clinical jargon, and shared decision making. AI could also explore other dimensions such as the meaning of words and phrases using NLP, turn taking, tone, and style. Many technical difficulties would have to be overcome to assess many of these variables, but the field is making progress.

An actual application of ML to feedback, but not in mental health, is provided by Pardo et al. (2019) in a course for first-year engineering students. Instructors developed in advance a set of feedback messages for levels of interaction with learning resources. For example, different feedback messages were provided depending on whether the student barely looked at video, watched a major portion, watched the whole video, or watched it several times. An ML algorithm selected the appropriate message to send the student through either email or the virtual learning environment. Compared to earlier cohorts who did not receive the feedback, those who did were more satisfied with the course and had better performance on the midterm. I can see how such a protocol could be used in mental health services.

An indication of the work that needs to be done in becoming more specific about feedback is a study conducted by Hooke et al. (2017). They provide feedback to patients with and without a trajectory showing expected progress and found that patients preferred the feedback with the expected change over time. They found that these patients preferred to have normative feedback with which they could compare their own ideographic progress.

Two systematic reviews that focused on implementing routine outcome measurement (ROM) concluded that while ROM has been shown to produce positive results, how to best implement ROM remains to be determined by future research (Gual-Montolio et al. 2020; MacKrill and Sorensen 2019). The authors of both reviews note several interesting points but focus on these two: how to integrate measurement into clinical practice and how organizations support staff in this effort. They highlight the importance of developing a culture of feedback in organizations. Neither review includes any studies using AI. While they call for more research to move this field forward, I do not think there will be much change until either measurement feedback systems are required by funders or service delivery organizations are paid for providing such systems.

Probably the most advanced work in this area that includes ML is being done by Lutz and his colleagues (Lutz et al. 2019). They have developed a measurement feedback system that includes the use of ML to make predictions and to provide clinicians with clinical decision support tools. They are able to predict dropouts and assign support tools to clinicians that are specific to the problems their clients are exhibiting, based on the data they have collected. Lutz and his colleagues are currently evaluating the system to influence clinical outcomes in a prospective study. This comprehensive feedback system provided clinical support tools with recommendations based on identification of similar patients to the treatment group but not to the control group. They already have some very promising results using three different treatment strategies (W. Lutz, personal communication, September 11, 2019).

Solutions to the Problem of Insufficiency of Treatment Precision

Almost all the research in this area has been on prediction and not in actually testing whether precision treatments are in fact better than standard treatments in improving mental health outcomes. Even these predictive studies are on extant databases rather than data collected specially for use in AI algorithms. With a few exceptions to be discussed later, this is the state of the art. To establish the practical usefulness of AI, we need to move beyond prediction to show actual mental health improvements that have clinical and not just statistical significance. There are some scholars who are carefully considering how to improve methodology to achieve better predictions (e.g., Garb and Wood 2019). In addition, Zilcha-Mano (2019) has a very thoughtful paper that describes traditional statistical and machine learning approaches to trying to answer the core question of what treatments work best for which patients, as well as the more general question about why psychotherapy works at all.

NLP has been used to analyze unstructured or textual material for identifying suicidal ideation in a psychiatric research database. Precision of 92% for identification of suicide ideation and 83% for suicide attempts has been found using NLP (Fernandes et al. 2017). A meta-analysis of 365 studies of prediction of suicide using traditional methodologies found only slightly better than chance predictions and no improvement in accuracy in 50 years (Franklin et al. 2017b). Recent ML decision support aids using large-scale biological and other data have been useful in predicting responses to different drugs for depression (Dwyer et al. 2018). Triantafyllidis and Tsanas (2019) conducted a literature review of pragmatic evaluations of nonpharmacological applications of ML in real-life health interventions from January 2008 through November 2018, following PRISMA guidelines. They found only eight articles that met their criteria from 7317 citations screened. Three dealt with depression and the remainder with other health conditions. Six of the eight produced significantly positive results, but only three were RCTs. There has been little rigorous research to support AI in real-world contexts.

Accuracy of prediction is one of the putative advantages of AI. But the advantage of predicting outcomes is not as relevant if a client prematurely leaves treatment. Thus, predicting premature termination is one of the key goals of an AI approach. In a pilot study to test whether AI could be beneficial in predicting premature termination, Bohus et al. (2018) were not able to adequately predict dropouts using 15 different ML approaches with 1159 responses to the Borderline Symptom List 23 (BSL-23)**.** However, they obtained some success when they combined the questionnaire data with 218 personal diary questionnaires from 14 patients, although they note that the sample is too small to draw any strong conclusions. This pilot study illustrates the importance of what data goes into the data set as well as our lack of knowledge of the data requirements we need to have confidence in as we select the appropriate data.

Duwe and Kim (2017) compared 12 statistical methods including ML approaches on their accuracy in predicting recidivism among 22,772 offenders. They found the newer ML algorithms generally performing modestly better. Kessler et al. (2015) used data from 38 U.S. Army and Department of Defense administrative data systems to predict suicides of soldiers who were hospitalized for a psychiatric disorder ($N = 40{,}820$). Within one year of hospitalization, 68 (0.17%) of the soldiers committed suicide. They used a statistical prediction rule based on ML that resulted in a high validity AUC value of 0.84. Kessler and his colleagues have continued this important work, which was discussed earlier.

Another approach to prediction was taken by Pearson et al. (2018) in predicting depression symptoms after an 8-week internet depression reduction program using 238 participants. They used an elastic net and random forest ML ensemble (combination) and compared it to a simple linear autoregressive model. They found that the ensemble method predicted an additional 8% of the variance over the non-ML approach. The authors offer several good technical suggestions about how to avoid some common errors in using ML. Moreover, the ML approach allowed them to identify specific module dosages that were related to outcomes that would be more difficult to determine using standard statistical approaches (e.g., detecting nonlinear relationships without having to specify them in advance). However, not all attempts to use AI are successful. Pelham et al. (2020) compared logistic regression and five different ML approaches to typical sum-score approaches to identify boys in the fifth grade who would be repeatedly arrested. ML performed no better than simple logistic regression when appropriate cross-validation procedures were applied. The authors emphasize the importance of cross-validation in testing ML approaches. In contrast, a predictive study of 1027 people with first-episode psychosis used AI to successfully predict poor remission and recovery one year later based only on baseline data (Leighton et al. 2019). The model was cross validated on two independent samples. A comprehensive synthesis of the literature of 300 studies that used ML or big data to address a mental health problem illustrated the wide variety of uses that currently exist; however, most dealt with detection and diagnosis (Shatte et al. 2019).

A critical view of the way psychiatry is practiced for the treatment of depression and how AI can improve that practice is provided by Tan et al. (2019). They note that most depression is treated with an “educated-guess-and-check approach in which clinicians prescribe one of the numerous approved therapies for depression in a stepwise manner” (p. 43). They posit that AI and especially deep learning have the ability to model the heterogeneity of outcomes and complexity of psychiatric disorders through the use large data sets. At this point, the authors have not provided any completed studies that have used AI, but two of the authors are shareholders in a medical technology company that is developing applications using deep learning in psychiatry. We are beginning to see commercial startups take an interest in mental health services even though the general health market is considerably bigger. Entrepreneurially motivated research may be important for the future of AI growth in mental health services, with traditional federal research grants to support this important developmental work, including such mechanisms as the Small Business Innovation Research (SBIR) program and the R21 and R34 NIH funding mechanisms.

One of the few studies that go beyond just prediction and actually attempt to develop a personalized treatment was conducted by Fisher et al. (2019). In a proof of concept study, the authors used Fisher's modular model of cognitive-behavioral therapy (CBT) and algorithms to develop and implement person-by-person treatments for anxiety and mood disorders for 32 adults. The participants were asked to complete surveys four times a day for about 30 days. The average improvement was better than found in comparison benchmark studies. The authors state that this is the first study to use pre-therapy multivariate time series data to generate prospective treatment plans.

Rosenfeld et al. (2019) describe several treatment delivery approaches that utilize AI. Woebot, for example, is a commercial product to provide CBT-based treatment using AI. The clients interact with Woebot through instant messaging that is later reviewed by a psychologist. It has been shown to have short-term effectiveness in reducing PHQ-9 scores of college students who reported depression and anxiety symptoms. The authors are optimistic that approaches like the ones described will lead to more widely available and efficacious treatment modalities. Applications of ML to addiction studies was the focus of a systematic review by Mak et al. (2019). They did an extensive search of the literature until December 2018 and could find only 17 articles. None of the studies involved evaluating a treatment.

I want to distinguish between the use of computer-assisted therapy, especially that provided through mobile apps, and the use of AI. In a review of these digital approaches to providing CBT for depression and anxiety, Wright et al. (2019) point out while many of these apps have been shown to be better than no treatment, they usually do not use AI to personalize them. Thus, they are less relevant to this paper and are not discussed in depth.

Ecological Momentary Interventions

Ecological momentary interventions (EMIs) are treatments provided to patients between sessions during their everyday lives (i.e., in real time) and in natural settings (Mohr et al. 2017). These interventions extend some aspects of psychotherapy to patients' daily lives to encourage activities and skill building in diverse conditions.

In the only systematic review available of EMIs, Colombo et al. (2019) found only eight studies that used EMIs to treat major depression, with only four different interventions. The common factor of these four interventions is that they provide treatment in real-time and are not dependent on planned sessions with a clinician. The authors report that participants were generally satisfied with the interventions, but there was variability in compliance and dropout rates among the programs. With only two studies that tested for effectiveness with RCTs, there is clearly a need for more rigorous evaluations.

Momentary reminders are typically used for behaviors such as medication adherence and management of symptoms. The more complex EMIs use algorithms to optimize and personalize systems. They also can use algorithms that changes the likelihood of the presentation of a particular intervention over time, based on past proximal outcomes. Schueller et al. (2017) note that EMIs are becoming more popular as a result of technological advances. These authors suggest the use of micro-randomized trials (MRTs) to evaluate them. An MRT uses a sequential factorial design that randomly assigns an intervention component to each person at multiple randomly chosen times. Each person is thus randomized many times. This complex design represents the dynamic nature of these interventions and how their outcomes correspond to different contextual features. AI is often used to develop algorithms to optimize and personalize the MRT over time. One interesting algorithm, called a "bandit algorithm," changes the intervention presented based on a past proximal outcome. As an example, Schueller et al. describe a hypothetical study to reduce anxiety through two different techniques—deep breathing and progressive muscle relaxation. The bandit algorithm may start the presentation of each technique with equal frequency but then shift more to the one that appears to be most successful for that individual. Thus, each treatment (a combination of deep breathing and progressive muscle relaxation) would be different for each person. Unlike RCTs, this method does not use group-level outcomes of average effect sizes but uses individual-level data. In the future, we might have personal digital mental health "therapists" or assistants that can deliver individualized combinations of treatments based on algorithms developed with AI that are data driven. Of course, this approach is best suited for these momentary interventions and would be difficult if not impossible to successfully apply to traditional treatment.

The Causality Conundrum: Do We Still Need RCTs?

I consider explicating the relationship between AI and causality to be a key factor in understanding whether AI is to be seen as replacing or as supplementing RCTs. Toward that end, I first consider whether observational data can replace RCTs using AI. Second, should a replacement not seem currently feasible, I explore ways to design studies that combine AI and RCTs to evaluate whether the AI approach produces better outcomes than non-AI enhanced interventions.

Observational Data and Causality

The journal *Prevention Science* devoted a special section of an issue to new approaches for making causal inferences from observational data (Wiedermann et al. 2019). An example is the paper by Shimizu (2019) that demonstrates the use of non-Gaussian analysis tools to infer causation from observational data under certain assumptions. Malinsky and Danks (2018) provide an extended discussion of the use of causal discovery algorithms to learn causal structure from observational data. In a similar fashion, Blöbaum et al. (2019) present a case for inferring causal direction between two variables by comparing the least-squares errors of prediction in both possible directions. Using data that meet some assumptions, they provide an algorithm that requires only a regression in both causal directions and a comparison of the least-square errors. Lechner's (2018) paper focuses on identifying the heterogeneity of treatment effects at the finest possible level or identifying what he calls groups of winners and losers who receive some treatment.

Hassani et al. (2018) hope to build a connection between researchers who use big data analysis and data mining techniques and those who are interested in causality analysis. They provide a guide that describes data mining applications in causality analysis. These include entity extractions, cluster analysis, association rule, and classification techniques. The authors also provide references to studies that use these techniques, key software, substantive areas in which they have been used, and the purpose of the applications. This is another bit of evidence that the issue of causality is being taken seriously and that some progress is being made. However, because of the newness of these publications, there is a lag in publications that are critical of these approaches; for example, D'Amour (2019) provides a technical discussion about why some approaches will not work but also suggests that others may be potentially effective. Clearly, caution is still warranted in drawing causal conclusion from observational data.

Chen (2019) provides a very interesting discussion of AI and causality but not from the perspective of the RCT issue that I raise here but as a much broader but still relevant point of view. He advances the key question about whether AI technology should be adopted in the medical field. Chen argues that there are two major deficits in AI, namely the causality deficit and the care deficit. The causality deficit refers to the inferior ability of AI to make accurate casual inferences, such as diagnosis, compared to humans. The care deficit is the comparative lack of ability of AI to care for a patient. Both deficits are interesting, but the one most germane to this paper is the causality deficit. Chen notes that AI represents statistical and not causal reasoning machines. He argues that AI is deficient compared to humans in causal reasoning, and, moreover, he doubts that there is a feasible way to deal with this lack of comparability in reasoning. He believes that AI is a model-blind approach in contrast to a human's more model-based approach to causal reasoning. Thus, causation for Chen is not an issue of experimental methodology (he never mentions RCTs in his paper), but a characteristic associated with humans and not computers. Chen does recognize that AI researchers are attempting to deal with the causality issue, for example, by briefly describing Pearl's (2000) directed acyclic graphs and nonparametric structural equation models. But Chen is skeptical that either the causality or care deficits will be overcome. He concludes that AI is best thought of as assisting humans in medical care and not replacing them. The relationship between AI and humans is a major concern of this paper.

Caliebe et al. (2019) see big data, and I would assume AI, as contributing to hypotheses generation that could then be tested in RCTs. The critical issues they see are related to the quality and quantity of big data. They quote an Institute of Medicine (IOM) report that refers to the use of big data and AI in medicine as "Learning Healthcare Systems" and states that these systems will "transform the way evidence on clinical effectiveness is generated and used to improve health and health care" (Institute of Medicine 2007, p. 1). Moreover, in 2007, the IOM suggested that alternative research methodologies will be needed. They do not acknowledge the conundrum that I have raised here; moreover, they do not see any need to consider changing any of our methodology or analyses. I have found many individual papers that describe how to solve the causality problem with AI (e.g., Kuang et al. 2020; Pearl 2019). Although these papers are complex, their mere existence gives me hope that this problem is being seriously considered.

In addition to the statistical and validity issues in trying to replace RCTs with observational data, there is the feasibility question. Although the data studied in much of the research reported in this paper are in the medical domain and deal primarily with medications, the characteristics of these data have some important lessons for mental health services. Bartlett et al. (2019) identified 220 trials published in the top seven highest impact medical journals. They then determined whether the intervention, medical condition, inclusion and exclusion criteria, and primary end points could be routinely obtained from insurance claims and/or electronic health data (EHR) data. These data are recognized by the FDA as what they term *real-world evidence*. They found that only 15% of the U.S.-based clinical trials published in high-impact journals in 2017 could be feasibly replicated through analysis of administrative claims or EHR data. The results suggest that potential for real-world evidence to replace clinical trials is very limited. At best, we can hope that they can complement trials. Given the paucity of data collected in mental health settings, the odds are that such data are even less available. Suggestions for improving the utility of real-world data for use in research are provided in an earlier article by some of these authors (Dhruva et al. 2018).

Pearl (2019) posits causal information in terms of the types of questions that, in his three-level model, each level answers. His first level is association; the second, intervention; and the third, counterfactual. Association is simply the statistical relationship or correlation. There is no causal information at this first level. The higher order levels can answer questions about the lower levels but not the other way around. Counterfactuals are the control groups in RCTs. They represent what would have happened if there had been no intervention. To Pearl, this unidirectional hierarchy explains why ML, based on associations, cannot provide causal statements like RCTs, which are based on counterfactuals. However, as noted earlier, Pearl does present an approach using what he calls *structural causal models* to "extract" causal relationships from associations. Pearl describes seven "talks" and accompanying tools that are accomplished in the framework provided by the structural causal models that are necessary to move from the lower levels to the counterfactual level to allow causal inferences. I would anticipate that there will be direct comparisons between this approach to causality and the randomized experiments like those done in program evaluation (Bickman and Reich 2014; Boruch et al. 2017).

Theory development or testing is usually not thought of as a strength of AI; instead, its lack of transparency, that is, the lack of explanatory power that would enable us to identify models/mechanisms that underlie outcomes, is seen as a major weakness. Coutanche and Hallion (2020) present a case for using feature ablation to test theories. This technique involves the removal or ablation of features from algorithms that have been thought to be theoretically meaningful and then seeing if there is a significant reduction in the predictive accuracy of the model. They have also studied whether the use of a different data set affects the predictive accuracy of a previously tested model in theoretically useful ways. They present a very useful hypothetical application of their approach to test theories using AI.

Can AI Replace RCTs?

It is clear that AI can be very useful in making predictions, but can it replace RCTs? Can AI perform the major function of RCTs, that of determining causality? The dependence on RCTs was one of the major limitations I saw as hindering the progress of mental health services research. While RCTs have their flaws, they are still considered by most as the best method for determining causal relationships. Is AI limited to being a precursor in identifying those variables that are good candidates for RCTs because they have high predictive values? The core conceptual problem is that while it is possible to compare two different but theoretically equivalent groups, one receiving the experimental treatment and the other the control condition, it is not possible to compare the same individuals on both receiving and not receiving the experimental treatment.

RCTs produce average effect sizes, but the ultimate purpose of precision mental health is to predict individualized effects. How do we reconcile these two very different aims? One approach is to use AI to identify the most predictive variables and then test them in a randomized experiment. Let us take a group of patients with the same disorder or problem. There may be several alternative treatments, but the most basic concept is to compare two conditions. In one condition, call it the traditional treatment condition in the RCT, everyone in that condition gets the same treatment. It is not individualized. In the second condition, call it the AI condition, everyone gets a treatment that is based on prior AI research. The latter may differ among individuals in dosage, timing, type of treatment, and so on. The simplest is medication that differs in dosage. However, a more nuanced design is a yoked design used primarily in operant and classical conditioning research. There have been limitations associated with this design, but these problems apply to conditioning research and not the application considered here (Church 1964).

To separate the effects of the individualization from the differences in treatment, I suggest using a yoked design. In this design, individuals who would be eligible to be treated with either the standard treatment or the AI-selected treatment would be yoked, that is, paired. Which participant of the pair received which condition would be randomized. First, the eligible participants would be randomly divided into two groups. The individuals in the AI group would get a treatment that was precisely designed for each person in that group, while those in the yoked control group would not; instead, those in the control group would receive the treatment that had been designed for his or her partner in the AI group. In this way, each participant would receive the same treatment, but only the AI group participants would be receiving individualized treatment. If the AI approach is superior, we would expect those in the AI group to have a superior average treatment effect compared to the control group, who received a treatment matched not to their individual characteristics but to those in the AI group.

We could also use an additional control group where the treatment is selected by a clinician. While this design would not easily identify which characteristics were responsible for its success, it would demonstrate whether individualized AI-based treatment was the causal factor. That is, we could learn that on the average, a precision approach is more effective than a traditional approach, but we would not be able to identify from this RCT which particular combination of characteristics made it more effective.

Of note is that the statistical power of this design would depend on the differences among the participants at baseline. For example, if the individuals were identical on measured covariates, then they would get the same personalized treatment, which practically would produce no useful information. Instead of yoking participants based on randomly assigning them as in the above example, we could yoke them on dissimilarity and then randomly assign each individual in the pair to AI-based treatment or a control condition that could be the same AI treatment or a clinician-assigned treatment. However, interesting this would be from a methodical point of view, I think this would also bring up ethical issues that are discussed next.

Of course, as with any RCT, there are ethical issues to consider. In many RCTs, the control group may receive standard treatment, which should not present any unusual ethical issues. However, in a yoked design, the control group participants will receive a treatment that was not selected for them on the basis of their characteristics. Moreover, the yoked design would make the formulation of the informed consent document problematic because it would have to indicate that participants in the control group would receive a treatment designed for someone else. One principle that should be kept in mind is equipoise: There should be consensus among clinicians and researchers that the treatments, a priori, are equivalent. In a yoked design, we must be assured that none of individualized treatments would harm the yoked control group members, and moreover, that there is no uniform agreement that the individualized treatment would be better for the recipient. That is, the research is designed to answer a question about relative effectiveness for which we do not know the answer.

A Third Path to Influencing Mental Health: The Role of Inflammation

Almost all of the research previously cited in this paper has dealt with psychosocial interventions, along with some research on interventions with medications. Clearly these are the two main approaches taken in providing services for mental health problems. However, in the last decade, a new approach to understanding mental illness has emerged from the field of psychoneuroimmunology. This relatively new field integrates research on psychology, neuroscience, and immunology to understand how these processes influence each other and, in turn, human health and behavior (Slavich 2019). I want to explore this relatively new approach to understanding mental health because I believe that it is a potentially rich field in which to apply AI.

Slavich and Irwin (2014) have combined diverse areas to show how stressors affect neural, physiologic, molecular, and genomic and epigenetic processes that mediate depression. They labeled this integrative theory the *social signal transduction theory of depression.* In a recent extension of this work, Slavich (2020) proposed social safety theory, which describes how social-environmental stressors that degrade experiences of social safety—such as social isolation and rejection—affect neural, immunologic, and genomic processes that increase inflammation and damage health.

A key aspect of this perspective is the role of inflammatory cytokines as key mediators of the inflammatory response (Slavich 2020). Cytokines are the biological endpoint of immune system activity and are typically measured in biobehavioral studies of stress and health. Cytokines promote the production of C-reactive protein, which is an inflammatory mediator like cytokines, but which also is a biomarker of inflammation that is assessed with a blood test. Cytokines also interact with the central nervous system and produce what have been labeled “sickness behaviors,” which include increased pain and threat sensitivity, anhedonia, fatigue, and social-behavioral withdrawal. While the relationship between inflammation and depression is well-established in adults, a systematic review and meta-analysis of studies with children and adolescents concluded that because of the small number of studies, more evidence was needed before drawing a similar conclusion for youth (D’Acunto et al. 2019). In contrast, a major longitudinal study of more than 4600 adults followed over 20 years found that participants who had stable high C-reactive protein levels were more likely to report clinically significant late-life depression symptoms (Sonsin-Diaz et al. 2020).

Chronic inflammation has been shown to be present in many psychiatric disorders including depression, schizophrenia, and PTSD, as well as in many other somatic and physical disease conditions (Furman et al. 2019). Chronic inflammatory diseases have been shown to be a major cause of death. A typical inflammatory response occurs when a threat is present and then goes away when there is no longer a threat. However, when the threat is chronic and unresolved, systemic chronic inflammation can occur and is distinct from acute inflammation. Chronic inflammation can cause significant damage to tissues and organs and break down the immune system tolerance.

What is especially interesting from a behavioral health perspective is that inflammatory activity can apparently be initiated by any psychological stressor, real or imagined. Thus, social and psychological stressors such as negative interpersonal relationships with friends and family, as well as physical stressors, can produce inflammation, which leads to increased risk of mental and physical health problems. This inflammatory response initially can have positive effects in that it can help increase survival in the short term, but it can also lead to a dysfunctional hypervigilance and anxiety that increases the risk of serious mental illness if chronic. The "cytokine storm" experienced by many COVID-19 patients is an example of the damage an uncontrolled immune response can cause (Konig et al. 2020). Although we do not know a great deal about how this process operates, it is clear that there is a strong linkage between inflammatory responses and mental disorders such as depression.

The role of the immune system in disease, especially brain inflammation related to brain microglial cells (i.e., neuroinflammation), is also receiving attention in the popular press (Nakazawa 2020). Psychoneuroimmunology research has explicated the linkage between the brain and the immune system, showing how stress affects the immune system, and how these interactions relate to mental illness. The relationships between these constructs suggest interventions that can be used to improve mental health. But much research remains to be done to identify specific processes and effective interventions. Research will require multidisciplinary teams to produce personalized interventions guided by each patient's specific level of neuroinflammation and genetic profiles. This process will need to be monitored by continuous feedback that I believe will be made more feasible with the application of AI. At present, there are some existing interventions that appear to be aligned with this approach that are being explored. These include the following.

Medications

Three anti-inflammatory medications have been found to reduce depressive symptoms in well-designed RCTs. These agents include celecoxib, usually used for treating excessive inflammation and pain, and etanercept and infliximab, which are used to treat rheumatoid arthritis, psoriasis, and other inflammatory conditions (Slavich 2019). However, there has not been a great deal of research in this area, so caution is warranted. A recent well-designed RCT with depressed youth tested aspirin, rosuvastatin (a statin), and a placebo and found no significant differences in depression symptoms (Berk et al. 2020).

Psychosocial Interventions

A meta-analysis explored the possible link between different types of psychosocial interventions, such as behavior therapy and CBT, and immune system function (Shields et al. 2020). The authors examined eight common psychosocial interventions, seven immune outcomes, and nine moderating factors in evaluating 56 RCTs. They found that psychosocial interventions were associated with a 19.1% improvement in good immune system function and a 4.1% decrease in detrimental immune function, on average. Moreover, the effects lasted for at least 6 months and were consistent across age, sex, and intervention duration. The authors concluded that psychosocial interventions are a feasible approach for influencing the immune system.

Repetitive Transcranial Magnetic Stimulation

Repetitive transcranial magnetic stimulation (rTMS) has been found to be an effective treatment for several mental illnesses, especially treatment-resistant depression (Mutz et al. 2019; Somani and Kar 2019; Voigt et al. 2019). While the literature is not clear on how rTMS produces its effect (Noda et al. 2015; Peng et al. 2018), I was curious about its relationship to neuroinflammation. I could find little in the research literature that addressed the relationship between inflammation and rTMS; therefore, I conducted an informal survey of 17 rTMS researchers who have published rTMS research in peer-reviewed journals and asked them the following:

> I suspect that rTMS is related to inflammation but the only published research that I could find on that relationship was two studies dealing with rats. Are you aware of any other research on this relationship? In addition, do you know of anyone using AI to investigate rTMS?

I received replies from all but 2 of the 17 researchers. About half said they were aware of some research that linked rTMS to inflammation and supplied citations. In contrast, only 20% were aware of any research on rTMS and AI. The latter noted some research that used AI on EEGs to predict rTMS outcomes. A most informative response was from the author of a review article that dealt with several different nontraditional treatments including rTMS on the hypothalamic–pituitary–adrenal (HPA) axis and immune function in the form of cytokine production in depression (Perrin and Parianti 2020). The authors found 15 relevant human studies (9 studies using rTMS) but were unable to conduct the meta-analysis because of significant methodological variability among studies. But they concluded that non-convulsive neurostimulation has the potential to impact abnormal endocrine and immune signaling in depression. Moreover, given that there is more information available than on other neurostimulation techniques, the research suggests that rTMS appears to reduce cytokines. Finally, there is some support from animal models (rats) that rTMS can have an anti-inflammatory effect on the brain and reduce depression and anxiety (Tiana et al. 2020). Moreover, four published studies showed that the efficacy of rTMS for schizophrenics could be predicted Koutsouleris et al. (2018b). Three other studies were able to use ML and EEG to predict outcomes of rTMS treatment for depression (Bailey et al. 2018; Hasanzadeh et al. 2019).

The existing literature indicates that metabolic activity and regional cerebral blood flow at the baseline can predict the response to rTMS in depression (Kar 2019). As these baseline parameters are linked to inflammation, it is worth studying responses to rTMS that predict inflammation. As noted by one of the respondents, "In summary, it is a relatively new field and there are no major multi-site machine learning studies in rTMS response prediction" (N. Koutsouleris, personal communication, March 15, 2020).

Finding Biomarkers

One of the significant limitations of measurement in mental health is the absence of robust biomarkers of inflammation. Furman et al. (2019) caution us that "Despite evidence linking SCI [systemic chronic inflammation] with disease risk and mortality, there are presently no standard biomarkers for indicating the presence of health-damaging chronic inflammation" (p. 1823). However, some biomarkers that are currently being explored for inflammation may be of some help. For example, Furman et al. (2019) are hopeful that a new approach using large numbers of inflammatory markers to identify predictors will produce useful information. A narrative review of inflammatory biomarkers for mood disorders was also cautious in drawing any conclusions from extant research because of "substantial complexities" (Chang and Chen 2020). It is also worth noting the emerging area of research on gut-brain communication and the relationship between microbiome bacteria and quality of life and mental health (Valles-Colomer et al. 2019). However, there is need for more research on the use of biomarkers.

The area of inflammation and mental health offers an additional pathway to uncovering the causes of mental illness but also, most importantly for this paper, potential services interventions beyond traditional medications and psychosocial interventions. Given the complexity, large number of variables from diverse data sets, and the emerging nature of this area, it appears that AI could be of great benefit in tying some potential biomarkers to effective interventions designed to produce better clinical outcomes. However, some caution is needed concerning the seemingly "hard data" provided by biomarkers. For example, Elliot et al. (2020) found in a meta-analysis of 90 experiments that one widely used biomarker, task-fMIR, had poor overall reliability and poor test–retest reliability in two other large studies. They concluded that these measures were not suitable for brain biomarker research or research on individual differences.

Problems and Limitations with AI

As noted in several places in this paper, AI is not without its problems and limitations. The next section of the paper discuses several of these problems.

Ethical and Legal Issues

AI may force the treatment developer to make explicit choices that are ethically ambiguous. For example, automobile manufacturers designing fully autonomous driving capabilities now have to be explicit about whose lives to value more in avoiding a collision—the driver and his or her passengers or a pedestrian. Should the car be programmed to avoid hitting a pedestrian, regardless of the circumstances, even if it results in the death of the driver? Mental health services do not typically have such clear-cut conflicts, but the need to weigh the potential side effects of a drug against potential benefits suggests that ethical issues will confront uses of AI in mental health.

Some research has shown that inherent bias in original data sets has produced biased (racist) decisions (Obermeyer et al. 2019; Veale and Binns 2017). An unresolved question is who has the responsibility for determining the accuracy and quality of original data set (Packin and Lev-Aretz 2018).

Data scientists operating with data provided by others may not have sufficient understanding of the complexity of the data to be sensitive to its limitations. Moreover, they may not consider it their responsibility to evaluate the accuracy of the data and attend to its limitations. Librenza-Garcia (2019) provides a comprehensive review of ethical issues in the use of large data sets with AI. The ethical issues in predicting major mental illness are discussed by Lawrie et al. (2019). They note that predictive algorithms are not sufficiently accurate at present, but they are progressing. The authors raise questions about whether people want to know their risk level for major psychiatric disorders, about individual and societal attitudes to such knowledge and the possible adverse effects of sharing such data, and about the possible impact of such information on early diagnosis and treatment. They urge conducting research in this area.

Related to the ethics issue but with more direct consequences to the health provider is the issue of legal responsibility in using an AI application. It is not clear what the legal liability is for interventions based on AI that go wrong. Who is responsible for such outcomes—the person applying the AI, the developer of the algorithm, or both? Price (2019) points out that providers typically do not have to be concerned about the legal liability of a negative outcome if they used standard care. Thus, if there are negative outcomes of some treatment but that treatment was the standard of care, there is usually no legal liability. However, currently AI is probably not seen as the standard of care in most situations. While this will hopefully change as evidence of the effectiveness of AI applications develops, currently the healthcare provider is at greater risk of legal liability in using an AI application that is different from the standard of care.

Weak Effect Sizes in Mental Health

I have previously discussed the insufficient evidence for the effectiveness of many of the interventions used in mental health services. This lack of strong evidence has implications for the use of AI in mental health services. In an insightful article on using AI for individual-level treatment predictions, Paulus and Thompson (2019) make several key observations and suggestions that are very relevant to the current paper. The authors summarize several meta-analyses of the weak evidence of effectiveness of mental health interventions and come to conclusions similar to those I have already stated. They also identify similar factors I have focused on in accounting for the modest effect sizes found in mental health RCTs. They point out that diagnostic categories are not useful if they are not aggregating homogenous populations. They suggest that what I call the diagnostic muddle may result from the nature of mental disorders themselves, for which there are many causes at many different levels, from the genetic to the environmental. Thus, there is no simple explanatory model. Paulus and Thompson note that prediction studies rarely account for more than a very small percentage of the variance. They recommend conducting large, multisite pragmatic RCTs that are clearly pre-defined with specific ML models and variables. Predictive models generated by this research then need to be validated with independent samples. This is a demanding agenda, but I think it is necessary if we are going to advance mental health services with the help of AI.

Lack of Transparency

Treatments are often considered black boxes that provide no understanding of how and why the treatment works (Kelley et al. 2010; Bickman 2008b). The problem of lack of transparency is compounded in the use of deep neural networks (Samek et al. 2017). At present we are not able to understand relationships between inputs and outcomes, because this AI technique does not adequately describe process. Deep neural networks may contain many hidden layers and millions of parameters (De Choudhury and Kikkoman 2018). However, this problem is now being widely discussed, and new technologies are being developed to make AI more transparent (Rauber et al. 2019; Kuang et al. 2020).

I do not believe it is possible to develop good theories of treatment effectiveness without this transparency. This is an important limitation of efforts to improve mental health services. But how important is this limitation? Early in my program evaluation career, I wrote about the importance of program theory (Bickman 1985, 1989). I argued that if individual studies were going to be conceptually useful, beyond local decisions such as program termination, then they must contribute to the broader goal of explaining why certain programs were effective and others not. This is in contrast to the worth and merit of a local program. A theory based evaluation of the program must add to our understanding of the theory underlying the program. While I still believe that generalizing to a broad theory of why certain interventions work is critical, at present it may be sufficient simply to increase the accuracy of our predictions, regardless of whether we understand why. As Stephens-Davidowitz (2017) argues, “in the prediction business, you just need to know that something works, not why” (p. 71). However, Turing Award winner Judea Pearl argued in his paper *Theoretical Impediments to Machine Learning with Seven Sparks from the Causal Revolution* (2018) that human-level AI cannot emerge from model-blind learning machines that ignore causal relationships.

One of the positive outcomes of the concern over transparency is the development of a subfield of AI that has been called *explainable artificial intelligence* (XAI). Adai and Berrada (2018) present a very readable description of this movement and show that it has been a growing area since 2016. They are optimistic that research in this area will go a long way toward solving the black box problem.

Need for Large Data Sets

Large data sets are required for some AI techniques, especially deep neural networks. While such data sets may be common in consumer behavior, social media, and hospital-based electronic health records, they are not common in community-based mental health services. The development and ownership of these data sets may be more important (and profitable) than ownership of specific AI applications. There is currently much turmoil over data ownership (Mittelstadt 2019). Ownership issues are especially important in the mental health field given the sensitivity of the data. In addition to the size and quality of the data set, longitudinal data are necessary for prediction. Collecting longitudinal data poses a particular problem for community-based services given the large treatment drop-out rate. In addition to the characteristics of the data, there is the need for competent data managers of large complex data sets.

The data requirements for mental health applications are more demanding than those for health in general. First, mental health studies usually do not involve the large samples that are found in general health. For example, the well-known Physicians' Health Study of aspirin to prevent myocardial infarction (MI) utilized more than 22,000 doctors in a RCT (Steering Committee of the Physicians' Health Study Research Group 1989). They found a reduction in MI that was highly statistically significant: $p < 0.00001$. The trial was stopped because it was thought that this was conclusive evidence that aspirin should be adopted for general prevention. However, the effect size was extremely small: a risk difference of 0.77% with $r^2 = 0.001$ (Sullivan and Feinn 2012). A study this size is not likely to occur in mental health. Moreover, such small effects would not be considered important even if they could be detected.

It is unlikely that very large clinical trials such as the aspirin study would ever be conducted in mental health. Thus, it is probable that data will have to be obtained from service data. But mental health services usually do not collect sufficiently fine-grained data from clients. While I was an early and strong proponent of what I called a measurement feedback system for services (Bickman 2008a), recent research shows that the collection of such data is rare in the real world. Until services start collecting these data as part of their routine services, it is unlikely that AI will have much growth with the limited availability of relevant data. There is, of course, a chicken and egg problem. A major reason why services do not collect data is the limited usefulness of data in improving clinical care. While AI may offer the best possibility of increasing the usefulness of regularly collected data, such data will not be available until policy makers, funders, and providers deem it useful and are willing to devote financial resources to such data collection analysis. At present, there are no financial incentives for mental health providers to collect such data even if they improved services.

Moustafa et al. (2018) made the interesting observation that psychology is behind other fields in using big data. AI and big data are not considered core topics in psychology. The authors suggest several reasons for this, including that psychology is mostly theory- and hypothesis-driven rather than data-driven, and that studies use small sample sizes and a small number of variables that are typically categorical and thus are not as amenable to AI. Moreover, most statistical packages used by psychologists are not well-equipped to analyze large data sets. However, the authors note that the method of clustering and thus differentiating among participants is used by psychologists and is in many ways similar to AI, especially deep neural networks, in trying to identify similar participants. Using ML methods such as random forest algorithms, the investigator can identify variables that best explain differences among groups or clusters. Instead of the typically few variables used by psychologists, AI can examine hundreds of variables.

As a note of caution, Rutledge et al. (2019) warn that "there is no silver bullet that can replace collecting enough data to generate stable and generalizable predictions" (p. 157). While there are techniques that are often used in low sample size situations (e.g., the elastic net and tree-based ensembles), researchers need replications with independent samples if they are to have sufficient confidence in their findings. Moreover, since big data are indeed big, they are easily misunderstood as automatically providing better results through smaller sampling errors. It is often not appreciated that the gain in precision drawn from larger samples may well be nullified by the introduction of additional population variance and biases.

Human Resources

Finding competent big data managers, data scientists, and programmers is a human resource problem.[2] In my experience, AI scientists who are able and want to collaborate with mental health services researchers are rare. Industry pays a lot more for these individuals than universities can afford. Moreover, even within the health field, mental health is a very small component of the cost of services, so it is often ignored in this area.

Difficulty and resistance are encountered in the implementation of new technologies. Clinicians are reluctant to adopt new approaches and to engage clients in new approaches and data collection procedures. Community mental health services have been slow to successfully adopt new technologies (Crutzen et al. 2014; Lattie et al. 2019; Yeager and Benight 2018). In their mixed methods study of community clinicians, Crutzen et al. (2014) found there were concerns about privacy, the wide range of therapeutic techniques used, disruptions in trust and alliance, managing crises, and organizational issues such as billing and regulations contained in the Privacy Rule established by the Health Insurance Portability and Accountability Act of 1996 (HIPAA) that inhibited the use of new technologies. Moreover, our current reimbursement policies do not support greater payment for better outcomes. Thus, there is little or no financial incentive for hard-pressed community services to improve their services at their own expense. In fact, I would argue that there is a disincentive to improve outcomes since it results in increased costs (at least initially), organizational disruption and potentially a loss of clients if it takes less time and effort to successfully treat them.

An interesting meta-issue has emerged from the widespread and ever-increasing investment in AI in healthcare. In a perceptive "Viewpoint" published in *JAMA*, Emanuel

[2] I would be happy to serve as a "matchmaker" for any AI programmers, data scientists (etc.), or behavioral scientists who are interested in collaborating on mental health projects. Just contact me describing your background and interests and I will try to put together like-minded researchers.

and Wachter (2019), argue that the major challenge facing healthcare is not that of obtaining data and new analytics but the achievement of behavior change among both clinicians and patients**.** They point out the major failures of Google and Microsoft in not recognizing the problems in translating evidence into practice in connection with their large, web-based repositories for storage of health records, Google Health and Microsoft HealthVault, both of which have been discontinued. They indicate that the long delays in translation are due not primarily to data issues or lack of accurate predictions, but to the absence of behavioral changes needed for adoption of these practices. For example, the collection of longitudinal data has been problematic. Another problem they note is that about half the people in the United States are nonadherent with medications. There is a huge gap between knowing what a problem is and actually solving it that "data gurus" seem to ignore. While this translation problem is evident in the sometimes narrow focus of AI promoters, it also represents an opportunity for the behavioral scientists engaged in AI research to marshal their skills and the knowledge gained from years of dealing with similar behavioral issues. The emergence of translational and implementation sciences, the latter more often led by behavioral scientists, can be of great service to the problems of applying AI to healthcare. The field of translational sciences has been developed and well-funded by the NIH in recognition of the difficulty in using (i.e., translating) laboratory studies into practice. In 2018, the budget for the Clinical and Translational Science Awards (CTSA) Program was over a half billion dollars from 2006 to 2020. However, as director of evaluation for Vanderbilt's Medical Center's CTSA program for many years, I became very familiar with the difficulties in applying medical research in the real world.

Complexity

Mental health is determined by multiple factors. It is unlikely that we will find a single vector such as a virus or a bacterium that causes mental illness. Thus, data demands can include multiple systems with biological, psychological, sociological, economic, and environmental factors. Within many of these domains, we do not have objective measures such as the lab tests found in medicine. Subjective self-reports are prone to many biases, and many of the symptoms are not observable by observers. The lack of a strong theory of mental disorders also makes it difficult to intelligently focus on only a few variables. Even with such apparently simple measures that include observations or recordings from multiple informants, we do not have a consensus on how to integrate them (Bickman et al. 2012a; Martel et al. 2017). However, I would expect that research generated with AI will contribute not only to improved treatment but also to enhanced theories by including heterogeneous clients and many data sources.

Trust and Confidentiality

Confidentiality and trust are key issues in mental health treatment. How will the introduction of AI affect the relationship between client and clinician? As noted earlier, there are problems, especially with deep learning, in interpreting the meaning of algorithmic solutions and predictions. Our ability to explain the algorithms to clients is problematic. While many research projects outside of mental health show that combining AI with human judgment produces the best outcomes, this research is still in its infancy.

Limited Use

A great deal has been written about AI in the context of medicine, but we need a reality check about the importance of AI in clinical practice. Ben-Israel et al. (2020) addressed the use of AI in a systematic review of the medical literature from 2000 to 2018. The authors focused on human studies that addressed a problem in clinical medicine using one or more forms of AI. Of the 386 studies, only 2% were prospective. None of the studies included a power analysis, and half did not report attrition data. Most were proof of concept studies. The authors concluded that their study showed that the use of AI in daily practice of clinical medicine is practically nonexistent. The authors acknowledge that use was defined by publication and that many applications of AI may be occurring without publication. Regardless, this study suggests that there are many barriers that must be overcome before AI is more widely used.

Implications of a Public Health Perspective

The self-help industry can provide perspective on digital apps, including some that use AI. It has been estimated that this sector was worth $9.9 billion in 2016 and is expected to be worth $13.2 billion in 2022 (La Rosa 2018). Part of that big dollar market is in digital mental health apps, although their precise monetary value is unknown.

More to the point is that we know little about the effectiveness of digital apps in the marketplace (Chandrashekar 2018). Moreover, many have warned that these unregulated and untested apps could be dangerous (Wykes 2019). In the United States, the publication of books is protected by the constitution, so there are no rules governing what can be published in the self-help sector. The market determines what gets accepted and used, regardless of effectiveness or negative side effects. But publication is limited by the cost of publishing and distribution. This is not the case for digital programs, where marginal costs of adding an additional user are negligible. Unlike other mental health interventions, there are no licensing or ethical standards governing their use. There are no data being uniformly collected on their use and their effects. Although there are U.S. government rules that can be applied to these apps (Armontrout et al. 2018), the law has many exceptions. The authors note that they could not find a single lawsuit related to software that diagnoses or treats a psychiatric condition. An interactive tool is provided by the Federal Trade Commission to help judge which federal laws might apply in developing an app (https://www.ftc.gov/tips-advice/business-center/guidance/mobile-health-apps-interactive-tool). It is clear that digital mental health apps will continue to grow. It is critical that services research and funding agencies do not overlook this development that might have potentially positive or negative effects.

Other Limitations of AI

These are but a few of the many areas or AI needing additional research and potential limitations to be addressed. An excellent discussion of these and other relates issues regarding the potential hype common in the AI field is provided in the National Academy of Medicine's monograph on the use of AI in healthcare (Matheny et al. 2019).

A thought-provoking paper by Hagendorff and Wezel (2019) classifies what AI can and cannot do. Some of the authors' concerns, such as measurement, completeness and quality of the data, and problems with transparency of algorithms, have already been discussed, so I will describe those that I feel are most relevant to mental health services. The authors describe two methodological challenges, the first being that the data used in AI systems are not representative of reality because of the way they are collected and processed. This can lead to biases and problems with generalizability. Second is the concern that supervised learning represents the past. Thus, prediction can be based only on the past and not on expectations of change; thus, in some respects, change is inhibited. Hagendorff and Wezel (2019) also note several societal challenges. One such challenge they cite is that many software engineers who develop these algorithms do not have sufficient knowledge of the sociological, psychological, ethical, and political consequences of their software. They suggest this leads to misinterpretations and misunderstandings about how the software will operate in society. The authors also note the scarcity of competent programmers. I noted earlier that this is especially the case in academia and particularly in the behavioral sciences. The authors highlight that AI systems often produce hidden costs. This includes hardware to run the AI systems and, I would add, the disruptive nature of the intrusion of AI into a workflow.

Among the technological challenges discussed by Hagendorff and Wezel, I believe the authors' focus on the big differences between human thinking and intelligent machines is especially relevant to mental health. Machines are in no way as complex as human brains; even AI's powerful neural networks, with more than a billion interconnections, represent only a tiny portion of the complexity of brain tissue. In order to obtain better convergence between machines and humans, Hagendorff and Wezel suggest that programmers follow the three suggestions made by Lake et al. (2016). First, programmers should move away from pattern recognition models, where most development started, to automated recognition of causal relationships. The second suggestion is to teach machines basic physical and psychological theories so that they have the appropriate background knowledge. The third suggestion is to teach machines to learn how to learn so that they can better deal with new situations.

The comparison between AI and human thought is the only aspect of their paper where Hagendorff and Wezel mention causality issues. They note the challenge related to the inflexibility of many algorithms, especially the supervised ones, where simply changing one aspect would result in processing errors because that aspect was not in the training data. Machines can be vastly superior to humans in some games where there are very specific inputs for achieving specific goals, but they cannot flexibly adapt to changes like humans. The authors suggest that promising technical solutions are being worked on to deal with this weakness in transferability. All these challenges will affect how well AI will work in mental health services. Most problems will probably be solved, but the authors believe that some of these challenges will never be met, such as dealing with the differences between human and computer cognition, which means that AI will never fully grasp the context of mental health services. The machine's construction of a person may lead to a fragmented or distorted self-concept that conflicts with the person's own sense of identity, which seems critical to any analysis of the person's mental health or lack thereof. I do not have a sense of how serious this and the other challenges will be for us in the future, but it is clear that there is a lot more we need to learn.

Yet another set of concerns, specifically about the variation in AI called deep learning (DL), was enumerated by Marcus (2018), an expert in DL. In a controversial paper in which he identified 10 limitations of DL, he noted that DL "may well be approaching a wall" (p. 3) where progress will slow or cease. For example, he noted that DL is primarily a statistical approach for classifying data, using neural networks with multiple layers. DL "maps" the relationships between inputs and outputs. While children may need only a few trials to correctly identify a picture of a dog, DL may need thousands or even millions of labeled examples before making correct identifications without the labels. Very large data sets are needed for DL. This is not the case for all ML techniques. I will not attempt to summarize the nine other limitations he sees with DL since many of them are noted elsewhere in this paper. He concludes that DL itself is not the problem; rather, the problem is that we do not fully understand the limitations of DL and what it does well. Marcus warns against excessive hype and unrealistic expectations. I am taking this advice personally, and I am not expecting my Tesla to be fully autonomous in 2020 as predicted by Elon Musk (Woodyard 2019).

Wolff (2018) provided an overview of how some of the problems of deep learning can be ameliorated. He responds to Marcus using many of the subheadings in Marcus's paper. He calls his framework the *SP theory of intelligence*, and its application is called the *SP computer model* (*SP* stands for simplicity and power). The theory was developed by Wolff to integrate observations and concepts across several fields including AI, computing, mathematics, and human perception and cognition, using information compression to unify them.

Advantages of AI

Despite these and other concerns previously described, I do think that the advantages of AI for moving mental health services forward outweigh its disadvantages. However, this summary of advantages does not attempt to balance in length or number the disadvantages described above. I do not think it is necessary to repeat the already described numerous applications and potential applications of AI that can be used to improve health services. Rather than repeating the numerous applications and potential applications of AI that can be used to improve health services, I highlight only a few key advantages.

One of the main advantages is the way AI deals with data. It can handle large amounts of data from diverse sources. This includes structured (quantitative) and unstructured (text, pictures, sound) data in the same analyses. Thus, it can integrate heterogeneous data from dissimilar sources. As noted earlier, the inclusion of non-traditional data such as those obtained from remote sensing (e.g., movement, facial expression, body temperature) will be responsible for a paradigm shift in what we consider relevant data.

AI, if widely adopted, has the potential to have a major impact on employment. While most of the popular press coverage has been on the potential negative effects of eliminating many jobs, there also are potential positive effects. AI can reduce the costs of many tasks, thus increasing productivity. On the human side, it can streamline routine work and eliminate many boring aspects of work. It thus can free up workers to engage in the more complex and interesting aspects of many jobs. Previous innovations have caused job dislocations. The classic loss of jobs in making buggy whips after the advent of automobiles is just one example. The inventions of the industrial age, such as steam engines, displaced many workers but also created many more new jobs. We know that many unskilled or semi-skilled jobs will be affected by AI in a major way. The elimination of cashiers with automated checkouts is now being implemented by Amazon. In these stores, you scan your phone, and then AI and cameras take over. You just put products in your bag or cart and leave when you are finished. Self-driving cars and trucks will greatly disrupt the transportation industry. We have weathered these disruptions in the past, but even the experts are unsure about how AI will influence jobs.

Probably the area in which there is the most positive potential in healthcare is when humans and machines collaborate in partnership. Here, AI augments human tasks but keeps humans in the center. Thus, physicians will no longer be separated by a laptop when speaking to a patient because AI will be able to record, take notes, and interpret the medical visit.

We have documented the shortage of mental health workers and the immense gap between mental health needs and our ability to fill them. Yes, we can train more clinicians, but our society seems unwilling to offer sufficient salaries to attract and keep such individuals. We have been experimenting with computers as therapists for more than 20 years, but now we finally have the technological resources to develop and implement such approaches. We have started to use chatbots to extend services, but in the near future, AI may allow us to replace the human therapist under some conditions (Hopp et al. 2018).

AI and the Future of Mental Health Services

In 1993, the computer scientist and science fiction author Vernor Vinge developed the concept of a singularity in which artificial intelligence would lead to a world in which robots attain self-consciousness and are capable of what are now human cognitive activities (Vinge 1993). Advocates and critics disagree on whether a singularity will be achieved and whether it would be a desirable development (Braga and Logan 2019). Braga and Logan, editors of a special issue of *Information* on the singularity and AI, conclude that although AI research is still in the early stage, the combination of human intelligence and AI will produce the best outcomes, but AI will never replace humans and we cannot fully depend on AI for the right answers. While these authors are well-informed, their crystal ball may not be clearer than anyone else's. The relevance of the singularity for healthcare lies in asking whether there will there be a time when AI-based computers are more effective and efficient than clinicians and will replace them. It is a question worth considering.

I have presented a comprehensive, wide-ranging paper dealing with AI and mental health services. I have described major deficiencies of our current services, namely the lack of sufficient access, inadequate implementation, and low efficiency/effectiveness. I summarized how precision medicine and AI have contributed to improving healthcare in general and how these approaches are being applied in precision psychiatry and mental health. The paper then describes research that shows how AI has been or can be used to help solve the five problems I noted earlier. I then described the disadvantages and advantages of AI. In reviewing all this information, I believe there is one factor that I have not discussed sufficiently that clearly differentiates the way mental health services have been delivered and the way I expect they will be delivered in the future. I want to focus this last section of the paper on what I believe is the most important and significant change that can occur. This change is reflected in a simple question: Is a human clinician necessary to deliver effective and efficient mental health services? I believe the answer to this question does not depend on the occurrence of the singularity but lies in the growth of AI research and its application to mental health services.

I think there is widespread agreement that there are significant problems with diagnoses and the quality of our measures. Moreover, most will probably agree that if AI can improve diagnoses and measures, then we should use utilize AI and let the results speak for themselves. The dependence on RCTs will probably not be resolved by AI research, but AI can clearly help inform what should be tested in RCTs. However, our current services overwhelmingly depend on human clinicians to deliver treatment. The problem with learning and feedback is that it requires clinicians to learn how to improve treatment over time with feedback. We are still uncertain about how well clinicians can learn from experience, training, and education (Bacon 2019). We also lack evidence of the best way to provide feedback to enhance that learning (Bickman 2008a; Dyason et al. 2020). The problem of treatment precision is also currently tied to having the clinician deliver the treatment. While we can expect AI to deliver more precise information about treatment planning, we still depend on the clinician to interpret and deliver it with fidelity with some evidence-based model. A precision approach requires the clinician to systematically deliver treatment that is most appropriate to a specific client. We do not have good evidence that most clinicians can do that.

The Critical Role of the Clinician

I believe no other issue generates a bigger emotional response than the idea of the changing the role of the clinician. No other issue has the economic impact on services as the position of the clinician. I believe this issue is the most critical to the future of mental health services and will be most affected by AI. I note that in 2016 in writing an introduction to an extensive special issue of this journal called "Therapist Effects in Mental Health Service Outcome" (King 2017), the authors of the introduction to that issue not did not note the potential role of AI in affecting clinicians (King and Bickman 2017). Change is happening rapidly.

Mental health services are not alone in facing the issue of the role of humans, although human clinicians are probably more central to the provision of mental health services than other health services. A similar issue of the role of humans in the provision of services is being played out in surgery. Surgery has been using robots for over 20 years (Bhandari et al. 2020), but the uptake has been slow for a variety of reasons. The next iteration of robot use is a move from using robots guided by surgeons to using robots assisted by AI and guided by surgeons. The use of AI may be seen as an intermediate step to fully autonomous AI-based robots not guided by surgeons. However, it is very clear that this progression is speculative and will take a long time to happen, if ever, given the consequences of errors. Closer to our everyday experience is the similar path that the development of autonomous driving involves as we move toward the point at which a human driver is no longer needed. Will mental health services follow a similar path?

Since we do not currently have a sufficient amount of research on using AI in treatment alone to inform us, we must look elsewhere for guidance. Two bodies of literature are relevant. One deals with the use of computers and other technologies that do not include the use of AI at present, the second with self-help in which the participation of the clinician is minimal or totally absent. First, let us consider the existing literature that contrasts technology-based treatments with traditional face-to-face psychotherapy. Then I will present some reviews of self-help research, followed by a description of the small amount of research using AI in treatment.

Technologically Based Interventions Not Using AI

A review of 25 studies of internet-delivered CBT (ICBT) to youth, using waitlist controls, supports the conclusion that CBT could be successfully adapted for internet-based treatment (Vigerlan et al. 2016). In a meta-analytic review of 9 meta-analyses, containing 166 studies of adult use of internet delivered via ICBT, the authors concluded that ICBT is as effective as face-to-face therapy (Andersson et al. 2019). Hermes et al. (2019) include websites, software, mobile aps, and sensors as instances of what they call *behavioral intervention technologies* (BIT). In their informative article, dealing primarily with implementation, they note that these technologies (they do not mention AI) can relate to a clinician in three ways:

(1) when intervention is delivered by the clinician and supported by BIT,
(2) when BIT provides the intervention with support from the clinician, or
(3) when intervention is fully automated with no role for the clinician.

This schema clearly applies to the AI interventions and the role of clinicians as well. Their conceptual model is helpful in understanding the parameters of implementation. They present a comprehensive plan for research to fill in the major gaps in the literature that addresses the question of comparative effectiveness of BIT and traditional treatment. Carlbring et al. (2018) conducted a systematic review and meta-analysis of 20 eligible studies of IBCT versus face-to-face CBT and reported that they produced equivalent outcomes, supporting the conclusions drawn by previous studies.

It is also important to consider the issue of therapeutic alliance (TA) and its relationship to internet-based treatment. TA, to a large extent, is designed to capture the human aspect of the relationship between the clinician and the client. There are thousands of correlational studies that have established that TA is a predictor of treatment outcomes (Flückiger et al. 2018); however, there are few studies of interventions that show a causal connection between TA and outcomes (e.g., Hartley et al. 2020). Moreover, the very nature of TA as trait-like or state-like, which is central to causal assumptions, is being questioned and is subject to new research approaches (Zilcha-Mano 2017) as well as to questions about how it should be measured (Bickman and Athay 2012) Regardless of my doubts about the importance of TA, the Fluckiger et al. (2018) meta-analysis found similar effect sizes ($r = 0.275$) for the alliance-outcome relationship in online interventions and in traditional face-to-face therapies. However, most of these studies were guided by a therapist, so the human factor was not totally absent. Penedo et al. (2020), in their study of a guided internet-based treatment, showed that it was important to align with the client's expectations and goals because these were related to outcomes, but no such relationship existed with the traditional third component of TA, bond with the supporting therapist, implying that TA might play a different role in internet-based treatments.

I was trained as a social psychologist and was a graduate student of Stanley Milgram (of the famous obedience experiments), so I was curious about the research on the relationship between technological virtual agents and humans beyond the context of mental health treatment. Several studies cited by Schneeberger et al. (2019) showed that robots could get people to do tiring, shameful, or deviant tasks. The authors found that participants obeyed these virtual agents similarly to the way they responded to humans in a video-chat format**.** The participants did the same number of shameful tasks regardless of who or what was ordering them. Moreover, doing the tasks produced the same level of shame and stress in the participant. They concluded that virtual agents and humans appear to have the same influence as human experimenters on participants. Of course, there are many limitations associated with generalizing from this laboratory study, which was conducted with female college students in Germany, but it does suggest that a great deal of research needs to be done on how humans relate to robots and virtual agents. Miner et al. (2019) suggest that use of conversational AI in psychotherapy can be an asset for improving access to care, but there is limited research on efficacy and safety.

Bibliotherapy and Self-Help

Can we learn about the role of the therapist from therapies that do not involve any therapist or technology? There is substantial research on self-help approaches from written material or what some call *bibliotherapy*. In general, research has supported the effectiveness of bibliotherapy before the advent of digital approaches. In 2010, Cuijpers et al. published a review of the literature that compared face-to face psychotherapy for depression and anxiety with guided self-help (i.e., with some therapist involvement) and concluded that they appeared comparable, but because there were so few studies in this comparison, this conclusion should be interpreted with caution. Has the situation changed in the last decade? In a comprehensive review and meta-analysis almost 10 years later, Bennett et al. (2019) conducted a review and meta-analysis of 50 studies. They concluded that self-help (both guided and unguided) had significant moderate to large effects on reducing symptoms of anxiety, depression, and disruptive behavior. However, there was also very high heterogeneity among the outcomes of these studies. Compared to face-to-face therapy, self-help was better than no treatment but slightly worse than face-to-face treatments, guided therapy was better than unguided, and computerized treatment was better than bibliographic treatment. It is important to note that none of the 50 studies were fully powered noninferiority trials, which would be a superior design. The authors concluded that their study showed potential near equivalence for self-help compared to face-to-face interventions, and their conclusions were consistent with several other reviews of self-help for mental health disorders in adults. The paper makes no mention of AI. Cuijpers et al. (2019) conducted a network meta-analysis of 155 trials of CBT addressing the question of whether format of delivery (individual, group, telephone-administered, guided self-help, or unguided self-help) influenced acceptability and effectiveness for these adult patients with acute depression. No statistically significant differences in effectiveness were found among these formats except that unguided self-help therapy was not more effective than care as usual but was more effective than a waitlist control group. The authors concluded that treatments using these different formats should be considered alternatives to therapist-delivered individual CBT. As in the previous publication, there was no mention of the use of AI, but Cuijpers believes that few if any of the studies reviewed in his publication used AI (P. Cuijpers, personal communication, March 24, 2020).

Technological Interventions Using AI

There is an emerging area of the use of AI in treatment that is informative. Tuerk et al. (2019), in a special section of *Current Psychiatry Reports* focusing on psychiatry in a digital age, describe several approaches to using technology in evidence-based treatments. Most relevant is their discussion of the use of AI in what has been called "conversational artificial intelligence" where there is a real-time interchange between a computer and a person. They note research that shows that this approach is low risk, high in consumer satisfaction, and high in self-disclosure. They suggest that there is a great deal of clinical potential in using AI in this manner. In a review of the literature from 1946 to 2018 on conversational agents used in the treatment of mental health problems, Gaffney et al. (2019) found only 13 qualifying studies out of an initial 30,853 with four being what they called full-scale RCTs. They concluded that the use of conversational agents was limited but growing. All studies showed reduced psychological distress, with the five controlled studies showing a significant reduction compared to control groups. However, the three studies that used active controls did not show significant differences between the waitlist controls and use of a conversational agent, although all showed improvement. The authors concluded that the use of conversational agents in therapy looks promising, but not surprisingly, more research is needed. A similar conclusion on conversational agents was reached in another independent review (Vaidyam et al. 2019). I have little doubt that more research will be forthcoming in this emerging area.

In summary, previous research using digital but not AI-powered ICBT, self-help (bibliotherapy), and AI-powered conversational agents suggests that effective treatment can be delivered without a human clinician under certain circumstances. I want to emphasize that these studies are suggestive but far from definitive. Rather, they suggest that the role of the clinician is worth more exploration, but they do not establish the conclusion that we do not need clinicians to deliver services. We need to know a great deal more about how AI-supported therapy operates in different contexts.

What Do Psychiatrists Think About Their Future and the Role of AI?

A survey of 791 psychiatrists from 22 countries asked about how technology will affect their future practice (Doraiswamy et al. 2020). Only 3.8% felt their jobs would become obsolete, and only a small minority (17%) felt that AI was likely to replace a human clinician in providing care. As much of the literature on the effects of AI on jobs suggests, those surveyed believed that AI would help in more routine tasks such as record keeping (75%) and synthesizing information, with about 50% believing their practices would be substantially changed. About 49% thought AI would have no influence or only minimal effect on their future work over the next 25 years. Another 47% thought their practices would be moderately changed by AI over the next 25 years. More than three quarters (83%) thought it unlikely that technology would ever be able to provide care as well as or better than the average psychiatrist. Only 30% of U.S.-based psychiatrists predicted that the potential benefits of future technologies or AI would outweigh the possible risks. Some of the specific tasks that psychiatrists typically perform, including mental status examination, evaluation of dangerous behavior, and the development of a personalized treatment plan, were also felt to be tasks that a future technology would be unlikely to perform as well. I do not think many psychiatrists in this study are prepared for the major changes in their practices that are highly likely to occur in the next quarter century.

In a thoughtful essay on the future of digital psychiatry, Hariman et al. (2019) draw a number of conclusions. They predict major changes in practice, with treatment by an individual psychiatrist alone becoming rare. Patients will receive treatment through their phones, participate in videoconferencing, and converse with chatbots. Clinicians will receive daily updates on the patients through remote sensing devices and self-report. AI will be involved in both diagnosis and treatment and will integrate diverse sources of information. Concerns over privacy and data security will increase. This is not the picture that the previously described survey of psychiatrists anticipated.

Brown et al. (2019) present the pros and cons of AI in an interesting debate format. On the pro side, the authors argue that while there are current limitations, the improvements in natural language processing (NLP) will lead to better clinical interviews. They point to research that shows people are more likely be honest with computers as a plus in obtaining more valid information from clients. They expect the AI "clinician" will be seen as competent and caring. They do note the danger that non-transparent AI will produce unintended negative side effects.

Those arguing against the use of AI clinicians acknowledge the technical superiority of AI to accomplish more routine tasks such as information gathering and tracking, but they point out the limitations even in the development of AI therapists. The lack of data needed to develop and test algorithms is critical. I have noted this in the discussion of the diagnostic muddle as a problem that AI can help solve, but these anti-AI authors argue that because psychiatrists disagree on diagnoses, there is no gold standard against which to measure the validity of AI models. This seems to be a rather unusual perspective from which to challenge change. They insightfully note that AI is different from human intelligence and does not perform well when presented with data that are different from training data. But the anti-AI authors acknowledge that more and better data may lead to improvement. Brown et al. (2019) argue that common sense is something that AI cannot draw on; however, this seems to be a weak argument when common sense has been demonstrated to be inaccurate under many situations. They conclude with the statement that psychiatry "will always be about connecting with another human to help that individual" (p. 2). This may be more wishful thinking than an accurate prediction about the future. Those arguing the pro position state that the "the advance of AI psychiatry is inexorable" (p. 3). On the other hand, the opponents of AI correctly point out that there is not yet sufficient evidence to draw a conclusion about the effectiveness of AI versus human clinicians. While there is disagreement about the potential advantages and disadvantages of AI, both sides agree that we need more and better research in this area.

Simon and Yarborough (2020) present the case that AI should not be a major concern for mental health. They argue that

> ideally, our field would abandon the term *artificial intelligence* in regard to actual diagnosis and treatment of mental health conditions. Using that term raises false hopes that machines will explain the mysteries of mental health and mental illness. It also raises false fears that all-knowing machines will displace human-centered mental health care. Big data and advanced statistical methods have and will continue to yield useful tools for mental health care. But calling those tools artificially intelligent is neither necessary nor helpful. (p. 220)

The authors further take the position that

> despite the buildup around artificial intelligence, we need not fear the imminent arrival of "The Singularity," that science fiction scenario of artificially intelligent computers linking together and ruling over all humanity. . . A scenario of autonomous machines selecting and delivering mental health treatments without human supervision or intervention remains in the realm of science fiction. (p. 220)

A more balanced approach to the role to the issue of replacement of clinicians by AI is presented by Ahuja (2019). After his review of the literature on medical specialists who may be replaced or more likely augmented by AI, his pithy take on this question is "Or, it might come to pass that physicians who use AI might replace physicians who are unable to do so" (Ahuja 2019, p. 19). Clearly, AI research will have to provide strong evidence of its effectiveness before AI will be accepted by some in the psychiatric community.

Conclusions

There are several pressing questions about how mental health services should be delivered and about the future of mental health services. Doubts about how much clinicians contribute to outcomes, our seeming inability to differentiate the effectiveness among clinicians except at the extremes, the lack of stability of employment of most community based clinicians, the poor track record on implementation of evidence-based programs, the cost of human services, the very limited availability of services especially where resources are inadequate—all lead to strong doubts about continuing the status quo of using clinicians as the primary way in which mental health services are delivered. In contrast, alternative approaches have many advantages. If scaled, AI therapists could be available to patients 24/7 and would not be bound to office hours. These AI therapists could represent any demographic or therapy style (e.g., directive) that the client preferred or that had been found to be more effective with a particular client. They can be specialists in any area for which there is sufficient research. In other words, not only can a personalized treatment plan be developed, but a personalized clinician (avatar) can be constructed for the best match with the client.

Of course, all these are putative advantages. As noted earlier, the application of AI is not without its risks and challenges, especially in putting together the interdisciplinary teams needed to accomplish this research. While I am optimistic about the potential contribution of AI to mental health services, it is just that—a potential. Extensive research will be needed to learn whether these approaches produce positive outcomes when compared to traditional face-to face treatment, while also dealing with the ethical issues raised by AI applications. Moreover, the quality of research needs significant improvement if we are going to have confidence in the findings. However, as exemplified by the rapid and uncontrolled growth of therapy apps, the world may not wait for rigorous supporting research before adopting a larger role for AI in mental health services.

While my brief summaries of findings of AI in the medical literature are supportive of the application of AI, I do not want to give the impression that these positive findings are accepted uncritically. A deeper reading of many of these studies exposes methodological flaws that temper enthusiasm. For example, in reviewing comparisons between healthcare professionals and deep learning algorithms in classifying diseases of all types using medical imaging, X. Liu et al. (2019a) conclude that the AI models are equivalent to the accuracy of healthcare professionals. This review is the first to compare the diagnostic accuracy of deep learning models to health-care professionals; however, only a small number of the 82 studies were direct comparisons. The authors also caution us by indicating what they labeled as the poor quality of many of the studies. The problems included low external validity (not done in a clinical practice setting), insufficient clarity in the reporting of results, lack of external validation, and lack of uniformity of metrics of diagnostic performance and deep learning terminology. However, the authors were encouraged by improvement in quality in the most recent studies analyzed. In commenting on the study, Cook (2019) noted other limitations and concluded that it is premature to draw conclusions about the comparative accuracy of AI versus human physicians. If we are not more cautious, she warns that we will experience "inflated expectations on the 2019 Gartner Hype Cycle" (p. E247). The latter refers to the examination of innovations and trends in AI. She cautions us to "stick to the facts, rather than risking a drop into the trough of disillusionment and a third major AI winter" (p. E247). Many issues are raised in Cook's paper, and the need to avoid the hype often found in the AI field is reiterated in the National Academy of Medicine's monograph on the use of AI in healthcare (Matheny et al. 2019).

Mental health services are changing. There are more than 10,000 mental health apps on the internet that are being used without much evidence of their effectiveness (Marshall et al. 2020; Bergin and Davis 2020; Gould et al. 2018). The explosion of mental health apps is the leading edge of future autonomous interventions. However, there is pressure to bring some order to this chaos. Probably the next innovation that will involve AI is its use in stepped therapy in which clients are typically triaged to low-intensity, low-cost care, monitored systematically, and stepped up to more intensive care if progress is not satisfactory (Mohr et al. 2019). In this schema, the low-cost care could be AI-based apps with little risk to the client. If more confidence is gained in the safety and effectiveness of this type of protocol, the use of AI-based treatment would be expected to increase.

Covid-19, Protests, and Mental Health Services

The Covid-19 pandemic will produce a major impact on mental health services. First, it is expected that the stresses caused by the pandemic will increase the demand for services (Qiu et al. 2020; Rajkumar 2020). Already poorly resourced mental health systems will not be able to meet this demand (Ćosić et al. 2020; Ho et al. 2020; Holmes et al. 2020), especially in low resourced countries. However, the biggest change will be in the service delivery infrastructure. Because of social distancing requirements, in-person delivery of therapy is being severely curtailed. While the major change at this time appears to be a shift to telemedicine (Shore et al. 2020; Van Daele et al. 2020), which is being adopted across almost all healthcare, there will need to be changes instituted in how clinicians are trained and supervised (Zhou et al. 2020). I have little doubt that AI will be adopted in order to increase efficiency and address the change in the service environment caused by the pandemic. In addition to changes initiated by the pandemic, there appear to be some changes in funding as a result of the protests concerning George Floyd's killing. There is reconsideration of shifting some funding from police services to mental health and conflict reduction services to be delivered by personnel outside law enforcement (Stockman and Eligon 2020). It will be difficult to meet this potential demand using the current infrastructure.

Integrating AI and the Mental Health Services Infrastructure

The literature on AI and medicine is replete with warnings about the difficulties we face in integrating AI into our healthcare system. As a program evaluator, I appreciate the position paper describing the urgent need for well-designed and competently conducted evaluations of AI interventions as well as the guidelines provided by Magrabi et al. (2019). More suggestions for improving the quality of research on supervised machine learning can be found in the paper by Cearns et al. (2019).

Celi et al. (2019) describe the future in a very brief essay that is worth quoting:

> Clinical practice should evolve as a hybrid enterprise with clinicians who know what to expect from, and how to work with, what is fundamentally a very sophisticated clinical support tool. Working together, humans and machines can address many of the decisional fragilities intrinsic to current practice. The human-driven scientific method can be powerfully augmented by computational methods sifting through

> the necessarily large amounts of longitudinal patient- and provider-generated data. (p. e256)

However, research on AI, data science, and other technologies is in its infancy if not the embryonic stage of development. I am fully immersed in the struggle to implement several types of technologies in practice. Changing the routine behavior of clinicians and clients is a major barrier to using new technologies, regardless of the effectiveness of these approaches. Emanuel and Wachter (2019) argue that the most important problem facing healthcare is not the absence of data or analytic approaches but turning predictions and findings into successful accomplishments through behavior change. Alongside the investment in technology and analytics, we need to support the research and applications of psychologists, behavioral economists, and those working in the relatively new field of translational and implementation research.

The emphasis on practical and implementable digital approaches requires a methodology that departs from the traditional efficacy approach, which does not focus on context and thus is difficult to translate to the real world. Mohr et al. (2019) suggest a solution-based approach that focuses on three stages that they label *create, trial* and *sustain.* Creation focuses on the initial stages of development, although not exclusively, and takes advantage of the unique characteristics of digital approaches that focus on engagement rather than trying to mimic traditional psychotherapy. Trial must be dynamic because digital technologies rapidly change; rapid evaluations are required, such as continuous quality improvement strategies (Bickman and Noser 1999). Sustainability requires more from investigators and evaluators than publication of results; they must also produce sustainable implementation that no longer depends on a research project for support.

AI Summers and Winters

We are currently in an AI summer in which there are important scientific breakthroughs and large investments in the application of AI (Hagendorff and Wezel 2019). But AI has had several winters when enthusiasm for AI has waned and unreasonable expectations have cooled. We were confronted with the reality that AI could not accomplish everything that people thought it could and that investors and journalists had hyped. AI, at least in the near term, will not be the superintelligence that will destroy humanity or the ultimate solution that will solve all problems. Enthusiasm for AI seems to run in cycles like the seasons. AI summers suffer from unrealistic expectations, but the winters bring an experience of disproportionate backlash and exaggerated disappointment. There was a severe winter in the late 1970s, and another in the 1980s and 1990s (Floridi 2020). Today, some are talking about another predictable winter (Nield 2019; Walch 2019; Schuchmann 2019). Floridi (2020) suggests that we can learn important principles from these cycles. First is whether AI is going to replace previous activities as the car did with the buggy, diversify activities as the car did with the bicycle, or complement and expand them as the plane did with the car. Floridi asks how acceptable an AI that survives another winter will be. He suggests that we need to avoid oversimplification and think deeply about with we are doing with AI. In the June 2020 issue of the Technology Quarterly of *The Economist* (2020), it is suggested that because AI's current summer is "warmer and brighter" than past ones because of widespread deployment of AI, "another full-blown winter is unlikely. But an autumnal breeze is picking up" (p. 4).

Looking Back and Forward

I have traced a path my career has taken from an almost exclusive focus on randomized experiments to consideration of the applications of AI. I have identified the main problems related to mental health services research's almost sole dependence on RCT methodology. I have linked the problems with this methodology with the lack of satisfactory progress in developing sufficiently effective mental health services. The recent availability of AI and the value now being placed on precision medicine have produced the early stages of a revolution in healthcare that will determine how treatment will be developed and delivered. I anticipate that in the very near future, a first-year graduate student will be contemplating the same questions that I raised 50 years ago, because they are still relevant, but this time he or she will realize that there are answers that were not available to me.

Acknowledgements This paper is part of a special issue of this journal titled "Festschrift for Leonard Bickman: The Future of Children's Mental Health Services." The issue includes a collection of 11 original children's mental health services research articles, this article, three invited commentaries on this article, and a compilation of letters in which colleagues reflect on my career and on their experiences with me. The word *Festschrift* is German and means a festival or celebration of the work of an author. There are many people to thank for their assistance in both the Festschrift and this paper. First, I want to acknowledge my two colleagues and friends, Nick Ialongo and Michael Lindsey, who spontaneously originated the idea of a Festschrift during a phone conversation with them. The folks at the Johns Hopkins Bloomberg School of Public Health were great in supporting the daylong event held on May 18, 2018. The many friends, family, former students and colleagues who traveled from around the country to attend and present made the event memorable. I am grateful to the committee that helped put this special issue together, which included Marc Atkins, Catherine Bradshaw, Susan Douglas, Nick Ialongo, Kim Hoagwood, and Sonja Schoenwald. This paper represents more than a yearlong effort for which many contributed including the scholars who provided email exchanges and ideas throughout the conceptualization and writing process. I thank the two editors of this special issue, Sonja and Catherine, who spent much of their valuable time on this project during a very difficult period. The manuscript was greatly improved through the efforts of my copy editor, Diana Axelsen. Most of all I thank Corinne Bickman, who has been my partner in life for almost 60 years and has managed this journal since its inception. Without her support and love none of this would have been possible.

Funding No external funding was used in the preparation or writing of this article.

Compliance with Ethical Standards

Conflict of interest From the editors: Leonard Bickman is editor-in-chief of this journal and thus could have a conflict of interest in how this manuscript was managed. However, the guest editors of this special issue, entitled "Festschrift for Leonard Bickman: The Future of Children's Mental Health Services," managed the review process. Three independent reviews of the manuscript were obtained and all recommended publication with some minor revisions, with which the editors concurred. While the reviewers were masked to the author, because of the nature of the manuscript is was not possible to mask the author for the reviewers. From the Author: The author reported receipt of compensation related to the Peabody Treatment Progress Battery from Vanderbilt University and a financial relationship with Care4 software. No other disclosures were reported.

Research Involving Human Subjects/Animals This article does not contain any research conducted by the authors involving human participants or animals.

Informed Consent This article did not involve any participants who required informed consent.

References

Abdullah, S., & Choudhury, T. (2018). Sensing technologies for monitoring serious mental illnesses. *IEEE Multimedia, 25*, 61–75.

Adai, A., & Berrada, M. (2018). Peeking inside the black-box: A survey on explainable artificial intelligence (XAI). *IEEE Access, 6*, 52138–52160. https://doi.org/10.1109/ACCESS.2018.2870052

Ahuja, A. (2019). The impact of artificial intelligence in medicine on the future role of the physician. *PeerJ, 7*, e7702. https://doi.org/10.7717/peerj.7702.

Alasuutari, P., Bickman, L., & Brannen, J. (Eds.). (2008). *The SAGE handbook of social research methods*. London: Sage.

Aliferis, C. F., Statnikov, A., Tsamardinos, I., Mani, S., & Koutsoukos, X. D. (2010). Local causal and Markov blanket induction for causal discovery and feature selection for classification part II: Analysis and extensions. *Journal of Machine Learning Research, 11*(7), 235–384.

Allsopp, K., Read, J., Corcoran, R., & Kinderman, P. (2019). Heterogeneity in psychiatric diagnostic classification. *Psychiatry Research, 279*, 15–22. https://doi.org/10.1016/j.psychres.2019.07.005.

Alves, P., Sales, C., & Ashworth, M. (2013). Enhancing the patient involvement in outcomes: A study protocol of personalised outcome measurement in the treatment of substance misuse. *BMC Psychiatry, 13*, 337–349. https://doi.org/10.1186/1471-244X-13-337.

Alves, P., Sales, C., & Ashworth, M. (2015). Personalising the evaluation of substance misuse treatment: A new approach to outcome measurement. *International Journal of Drug Policy, 26*, 333–335. https://doi.org/10.1016/j.drugpo.2014.11.014.

Anderson, T., McClintock, A. S., Himawan, L., Song, X., & Patterson, C. L. (2016). A prospective study of therapist facilitative interpersonal skills as a predictor of treatment outcome. *Journal of Consulting and Clinical Psychology, 84*(1), 57–66.

Andersson, G., Per Carlbring, P., Titov, N., & Lindefors, N. (2019). Internet interventions for adults with anxiety and mood disorders: A narrative umbrella review of recent meta-analyses. *The Canadian Journal of Psychiatry/La Revue Canadienne de Psychiatrie, 64*(7), 465–470. https://doi.org/10.1177/0706743719839381.

Armontrout, J. A., Torous, J., Cohen, M., McNiel, D. E., & Binder, R. (2018). Current regulation of mobile mental health applications. *Journal of the American Academy of Psychiatry and the Law, 46*(2), 204–211. https://doi.org/10.29158/JAAPL.003748-18.

August, G. J., & Gewirtz, A. (2019). Moving toward a precision-based, personalized framework for prevention science: Introduction to the special issue. *Prevention Science, 20*(1), 1–9. https://doi.org/10.1007/s11121-018-0955-9.

August, G. J., Piehler, T. F., & Blomquist, M. L. (2016). Being "SMART" about adolescent conduct problems prevention: Executing a SMART pilot study in a juvenile diversion agency. *Journal of Clinical Child and Adolescent Psychology, 45*(4), 495–509. https://doi.org/10.1080/15374416.2014.945212.

Bacon, S. A. (2019). Constructionist extension of the contextual model: Ritual, charisma, and client fit. *Journal of Psychotherapy Integration*. https://doi.org/10.1037/int0000188.

Bailey, N. W., Hoy, K. E., Rogasch, N. C., Thomson, R. H., McQueen, S., Elliot, D., et al. (2018). Responders to rTMS for depression show increased front to-midline theta and theta connectivity compared to non-responders. *Brain Stimulation, 11*(1), 190–203. https://doi.org/10.1016/j.brs.2017.10.015.

Bains, R. M., & Diallo, A. F. (2016). Mental health services in school-based health centers: Systematic review. *Journal of School Nursing, 32*(1), 8–19. https://doi.org/10.1177/1059840515590607.

Bakker, M. J., Greven, C. U., Buitelaar, J. K., & Glennon, J. C. (2017). Practitioner review: Psychological treatments for children and adolescents with conduct disorder problems: A systematic review and meta-analysis. *Journal of Child Psychology and Psychiatry and Allied Disciplines, 58*(1), 4–18. https://doi.org/10.1111/jcpp.12590.

Bartlett, V. L., Dhruva, S. S., Shah, N. D., Ryan, P., & Ross, J. S. (2019). Feasibility of using real-world data to replicate clinical trial evidence. *JAMA Network Open, 2*(10), e1912869. https://doi.org/10.1001/jamanetworkopen.2019.12869.

Behar, L. B. (1997). The Fort Bragg evaluation: A snapshot in time. *American Psychologist, 52*(5), 557–559. https://doi.org/10.1037/0003-066X.52.5.557.

Bell, I. H., Lim, M. H., Rossell, S. L., & Thomas, N. (2017). Ecological momentary assessment and intervention in the treatment of psychotic disorders: A systematic review. *Psychiatric Services, 68*(11), 1172–1181. https://doi.org/10.1176/appi.ps.201600523.

Bennett, S. D., Cuijpers, P., Ebert, D. D., Smith, M. M., Coughtrey, A. E., Heyman, I., et al. (2019). Practitioner review: Unguided and guided self-help interventions for common mental health disorders in children and adolescents: A systematic review and meta-analysis. *Journal of Child Psychology and Psychiatry*. https://doi.org/10.1111/jcpp.13010.

Ben-Israel, D., Bradley Jacobs, W., Casha, S., Lang, S., Ryu, W. H. A., de Lotbiniѐe-Bassett, M., & Cadotte, D. W. (2020). The impact of machine learning on patient care: A systematic review. *Artificial Intelligence in Medicine, 103*, 101785, ISSN 0933–3657. https://doi.org/10.1016/j.artmed.2019.101785.

Bergin, A., & Davis, E. B. (2020). Technology matters: Mental health apps-separating the wheat from the chaff. *Child and*

Adolescent Mental Health, 25(1), 51–53. https://doi.org/10.1111/camh.12363.

Berk, M., Mohebbi, M., Dean, O. M., Cotton, S. M., Chanen, A. M., Dodd, S., et al. (2020). Youth depression alleviation with anti-inflammatory agents (YoDA-A): A randomised clinical trial of rosuvastatin and aspirin. *BMC Medicine, 18*(16), 20. https://doi.org/10.1186/s12916-019-1475-6.

Berkel, C., Gallo, C. G., Sandler, I. N., Mauricio, A. M., Smith, J. D., & Brown, C. H. (2019). Redesigning implementation measurement for monitoring and quality improvement in community delivery settings. *Journal of Primary Prevention, 40*(1), 111–127. https://doi.org/10.1007/s10935-018-00534-z.

Bernecker, S. L., Coyne, A. E., Constantino, M. J., & Ravitz, P. (2017). For whom does interpersonal psychotherapy work? A systematic review. *Clinical Psychology Review, 56*, 82–93. https://doi.org/10.1016/j.cpr.2017.07.001.

Bhandari, M., Zeffiro, T., & Reddiboina, M. (2020). Artificial intelligence and robotic surgery: Current perspective and future directions. *Current Opinion in Urology, 30*(1), 48–54. https://doi.org/10.1097/MOU.0000000000000692

Bickman, L. (1972). Social influence and diffusion of responsibility in an emergency. *Journal of Experimental Social Psychology, 8*(5), 438–445. https://doi.org/10.1016/0022-1031(72)90069-8.

Bickman, L. (1974a). The social power of a uniform. *Journal of Applied Social Psychology, 4*(1), 47–61. https://doi.org/10.1111/j.1559-1816.1974.tb02599.x.

Bickman, L. (1974b). Social roles and uniforms: Clothes make the person. *Psychology Today, 7*(11), 48–51.

Bickman, L. (1985). Improving established statewide programs: A component theory of evaluation. *Evaluation Review, 9*(2), 189–208. https://doi.org/10.1177/0193841X8500900206.

Bickman, L. (1989). Barriers to the use of program theory. *Evaluation and Program Planning, 12*(4), 387–390. https://doi.org/10.1016/0149-7189(89)90056-6.

Bickman, L. (1995). The Fort Bragg demonstration project: A managed continuum of care. *The Child, Youth, and Family Services Quarterly, 18*(3), 2–5.

Bickman, L. (1996). A continuum of care: More is not always better. *American Psychologist, 51*(7), 689–701. https://doi.org/10.1037/0003-066X.51.7.689.

Bickman, L. (1997). Resolving issues raised by the Ft. Bragg findings: New directions for mental health services research. *American Psychologist, 52*, 562–565. https://doi.org/10.1037/0003-066X.52.5.562.

Bickman, L. (1999). Practice makes perfect and other myths about mental health services. *American Psychologist, 54*(11), 965–978. https://doi.org/10.1037/h0088206.

Bickman, L. (2006). My life as an applied social psychologist. *Current Psychology, 25*(2), 67–92. https://doi.org/10.1007/s12144-006-1005-5.

Bickman, L. (2008a). A measurement feedback system (MFS) is necessary to improve mental health outcomes. *Journal of the American Association of Child and Adolescent Psychiatry, 47*, 1114–1119. https://doi.org/10.1097/CHI.0b013e3181825af8.

Bickman, L. (2008b). Why don't we have effective mental health services? [Editorial]. *Administration and Policy in Mental Health and Mental Health Services Research, 35*(6), 437–439. https://doi.org/10.1007/s10488-008-0192-9.

Bickman, L. (2012). Why can't mental health services be more like modern baseball? *Administration and Policy in Mental Health and Mental Health Services Research, 39*(1–2), 1–2. https://doi.org/10.1007/s10488-012-0409-9.

Bickman, L., & Athay, M. I., (Eds.) (2012). Youth Mental Health Measurement (special issue) *Administration and Policy in Mental Health and Mental Health Services Research, 39,*1-2. ISSN: 0894-587X (Print) 1573-3289 (Online)

Bickman, L., Douglas, S., Vides de Andrade, A. R., Tomlinson, M., Gleacher, A., Olin, S., et al. (2015). Implementing a measurement feedback system: A tale of two sites. *Administration and Policy in Mental Health and Mental Health Services Research, 43*, 410–425. https://doi.org/10.1007/s10488-015-0647-8.

Bickman, L., Guthrie, P., Foster, E. W., Lambert, E. W., Summerfelt, W. T., Breda, C., et al. (1995). *Evaluating managed mental health care: The Fort Bragg experiment*. New York: Plenum.

Bickman, L., & Henchy, T. (Eds.). (1972). *Beyond the laboratory: Field research in social psychology*. New York: McGraw-Hill.

Bickman, L., Karver, M., & Schut, L. J. A. (1997a). Clinician reliability and accuracy in judging appropriate level of care. *Journal of Consulting and Clinical Psychology, 65*(3), 515–520. https://doi.org/10.1037/0022-006X.65.3.515.

Bickman, L., Kelley, S., & Athay, M. (2012a). The technology of measurement feedback systems. *Couple and Family Psychology: Research and Practice, 1*(4), 274–284. https://doi.org/10.1037/a0031022.

Bickman, L., Kelley, S. D., Breda, C., Vides de Andrade, A. R., & Riemer, M. (2011). Effects of routine feedback to clinicians on mental health outcomes of youths: Results of a randomized trial. *Psychiatric Services, 62*(12), 1423–1429. https://doi.org/10.1176/appi.ps.002052011.

Bickman, L., Lambert, E. W., Andrade, A. R., & Penaloza, R. (2000). The Fort Bragg Continuum of Care for children and adolescents: Mental health outcomes over five years. *Journal of Consulting and Clinical Psychology, 68*(4), 710–716. https://doi.org/10.1037/0022-006x.68.4.710.

Bickman, L., Lyons, A., & Wolpert, M. (2016). Achieving precision mental health through effective assessment, monitoring, and feedback processes. *Administration and Policy in Mental Health and Mental Health Services Research, 43*, 271–276. https://doi.org/10.1007/s10488-016-0718-5.

Bickman, L., & Noser, K. (1999). Meeting the challenges in the delivery of child and adolescent mental health services in the next millennium: The continuous quality improvement approach. *Applied and Preventive Psychology, 8*(4), 247–255. https://doi.org/10.1016/S0962-1849(05)80039-3.

Bickman, L., & Reich, S. (2014). Randomized controlled trials: A gold standard or gold plated? In S. I. Donaldson, C. A. Christie, & M. M. Mark (Eds.), *What counts as credible evidence in applied research and evaluation practice?*. Thousand Oaks, CA: Sage.

Bickman, L., & Rog, D. (Eds.). (2009). *The SAGE handbook of applied social research methods*. Thousand Oaks, CA: Sage.

Bickman, L., & Rog, D. (Eds.). (2016). *The evaluation handbook: An evaluator's companion*. New York: Guilford Press.

Bickman, L., & Rosenbaum, D. P. (1977). Crime reporting as a function of bystander encouragement, surveillance, and credibility. *Journal of Personality and Social Psychology, 35*(8), 577–586. https://doi.org/10.1037/0022-3514.35.8.577.

Bickman, L., Smith, C. M., Lambert, E. W., & Andrade, A. R. (2003). Evaluation of a congressionally mandated wraparound demonstration. *Journal of Child and Family Studies, 12*(2), 135–156. https://doi.org/10.1023/A:1022854614689.

Bickman, L., Summerfelt, W. T., & Noser, K. (1997b). Comparative outcomes of emotionally disturbed children and adolescents in a system of services and usual care. *Psychiatric Services, 48*(12), 1543–1548. https://doi.org/10.1176/ps.48.12.1543.

Bickman, L., Vides de Andrade, A. R., Athay, M. M., Chen, J. I., De Nadai, A. S., Jordan-Arthur, B., et al. (2012b). The relationship between change in therapeutic alliance ratings and improvement in youth symptom severity: Whose ratings matter the most? *Administration and Policy in Mental Health and Mental Health Services Research, 39*(1–2), 78–89. https://doi.org/10.1007/s10488-011-0398-0.

Bickman, L., Wighton, L. G., Lambert, E. W., Karver, M. S., & Steding, L. (2012c). Problems in using diagnosis in child and adolescent mental health services research. *Journal of Methods and Measurement in the Social Sciences, 3*(1), 1. https://doi.org/10.2458/v3i1.16110.

Blöbaum, P., Janzing, D., Washio, T., Shimizu, S., & Schölkopf, B. (2019). Analysis of cause-effect inference by comparing regression errors. *PeerJ Computer Science, 5*, e169. https://doi.org/10.7717/peerj-cs.169.

Bohus, M., Gimbel, S., Goerg, N., Humm, B. G., Schüller, M., Steffens, M., & Vonderlin, R. (2018). Improving machine learning prediction performance for premature termination of psychotherapy. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, *11089 LNAI* (517), 141–151. https://doi.org/10.1007/978-3-319-99344-7_13

Boruch, R., Yang, R., Hyatt, J., & Turner, H. (2017). Randomized controlled trials. In B. Greve (Ed.), *Handbook of social policy evaluation*. Chettenham, UK: Edward Elgar.

Boukhechba, M., Chow, P., Fua, K., Teachman, B. A., & Barnes, L. E. (2018). Predicting social anxiety from global positioning system traces of college students: Feasibility study. *Journal of Medical Internet Research, 5*(3), e10101. https://doi.org/10.2196/10101.

Braga, A., & Logan, R. K. (2019). AI and the singularity: A fallacy or a great opportunity? *Information, 10*(2), 73. https://doi.org/10.3390/info10020073.

Brattland, H., Koksvik, J. M., Burkeland, O., Gråwe, R. W., Klöckner, C., Ryum, T., et al. (2018). The effects of routine outcome monitoring (ROM) on therapy outcomes in the course of an implementation process: A randomized clinical trial. *Journal of Counseling Psychology, 65*(5), 641–652. https://doi.org/10.1037/cou0000286.

Brown, C., Story, G. W., Mourão-Miranda, J., & Baker, J. T. (2019). Will artificial intelligence eventually replace psychiatrists? *The British Journal of Psychiatry, 12*, 1–4. https://doi.org/10.1192/bjp.2019.245.

Bryant, D., & Bickman, L. (1996). Methodology for evaluating mental health case management. *Evaluation and Program Planning, 19*(2), 121–129. https://doi.org/10.1016/0149-7189(96)00003-1.

Buskirk, T. D., Kirchner, A., Eck, A., & Signorino, C. S. (2018). An introduction to machine learning methods for survey researchers. *Survey Practice, 11*(1), 1–10. https://doi.org/10.29115/sp-2018-0004.

Bzdok, D., & Karrer, T. (2018). Single-subject prediction: A statistical paradigm for precision psychiatry. HAL Id : hal-01714822.

Bzdok, D., & Meyer-Lindenberg, A. (2018). Machine learning for precision psychiatry: Opportunities and challenges. *Biological Psychiatry: Cognitive Neuroscience and Neuroimaging, 3*(3), 223–230. https://doi.org/10.1016/j.bpsc.2017.11.007.

Caliebe, A., Leverkus, F., Antes, G., & Krawczak, M. (2019). Does big data require a methodological change in medical research? *BMC Medical Research Methodology, 19*(1), 1–6. https://doi.org/10.1186/s12874-019-0774-0.

Carcone, A. I., Hasan, M., Alexander, G. L., Dong, M., Eggly, S., Hartlieb, K. B., et al. (2019). Developing machine learning models for behavioral coding. *Journal of Pediatric Psychology, 44*(3), 289–299. https://doi.org/10.1093/jpepsy/jsy113.

Carlbring, P., Andersson, C., Cuijpers, P., Riper, H., & Hedman-Lagerlöf, E. (2018). Internet-based vs. face-to-face cognitive behavior therapy for psychiatric and somatic disorders: An updated systematic review and meta-analysis. *Cognitive Behaviour Therapy, 47*(1), 1–18. https://doi.org/10.1080/16506073.2017.1401115.

Cearns, M., Hahn, T., & Baune, B. T. (2019). Recommendations and future directions for supervised machine learning in psychiatry. *Translational Psychiatry*. https://doi.org/10.1038/s41398-019-0607-2.

Celi, L. A., Fine, B., & Stone, D. J. (2019). An awakening in medicine: The partnership of humanity and intelligent machines. *The Lancet, Digital Health, 1*(6), e255–e257. https://doi.org/10.1016/S2589-7500(19)30127-X.

Chandrashekar, P. (2018). Do mental health mobile apps work: Evidence and recommendations for designing high-efficacy mental health mobile apps. *Mhealth*. https://doi.org/10.21037/mhealth.2018.03.02.

Chanfreau-Coffinier, C., Peredo, J., Russell, M. M., Yano, E. M., Hamilton, A. B., Lerner, B., et al. (2019). A logic model for precision medicine implementation informed by stakeholder views and implementation science. *Genetics in Medicine, 21*(5), 1139–1154. https://doi.org/10.1038/s41436-018-0315-y.

Chang, H. H., & Chen, P. S. (2020). Inflammatory biomarkers for mood disorders: A brief narrative review. *Current Pharmaceutical Design, 26*(2), 236–243. https://doi.org/10.2174/1381612826666200115100726.

Chen, M. (2019). A tale of two deficits: Causality and care in medical AI. *Philosophy & Technology*. https://doi.org/10.1007/s13347-019-00359-6.

Church, R. M. (1964). Systematic effect of random error in the yoked control design. *Psychological Bulletin, 62*(2), 122–131. https://doi.org/10.1037/h0042733.

Clarkson, T., Kang, E., Capriola-Hall, N., Lerner, M. D., Jarcho, J., & Prinstein, M. J. (2019). Meta-analysis of the RDoC social processing domain across units of analysis in children and adolescents. *Journal of Clinical Child & Adolescent Psychology*. https://doi.org/10.1080/15374416.2019.1678167.

Cohen, A. S. (2019). Advancing ambulatory biobehavioral technologies beyond "proof of concept": Introduction to the special section. *Psychological Assessment, 31*(3), 277–284. https://doi.org/10.1037/pas0000694.

Colombo, D., Fernández-Álvarez, J., Patané, A., Semonella, M., Kwiatkowska, M., Garcia-Palacios, A., et al. (2019). Current state and future directions of technology-based, ecological momentary assessment and intervention for major depressive disorder: A systematic review. *Journal of Clinical Medicine, 8*(4), 465. https://doi.org/10.3390/jcm8040465.

Connolly Gibbons, M. B., Kurtz, J. E., Thompson, D. L., Mack, R. A., Lee, J. K., Rothbard, A., et al. (2015). The effectiveness of clinician feedback in the treatment of depression in the community mental health system. *Journal of Consulting and Clinical Psychology, 83*(4), 748–759. https://doi.org/10.1037/a0039302.

Cook, C. R., Kilgus, S. P., & Burns, M. K. (2018). Advancing the science and practice of precision education to enhance student outcomes. *Journal of School Psychology, 66*, 4–10. https://doi.org/10.1016/j.jsp.2017.11.004.

Cook, T. S. (2019). Human versus machine in medicine: Can scientific literature answer the question? *The Lancet: Digital Health, 1*(6), e246–e247.

Ćosić, K., Popović, S., Šarlija, M., & Kesedžić, I. (2020). Impact of human disasters and COVID-19 pandemic on mental health: Potential of digital psychiatry. *Psychiatria Danubina, 32*(1), 25–31. https://doi.org/10.24869/psyd.2020.25.

Costello, E. J., He, J. P., Sampson, N. A., Kessler, R. C., & Merikangas, K. R. (2014). Services for adolescents with psychiatric disorders: 12-month data from the National Comorbidity Survey-Adolescent. *Psychiatric Services, 65*(3), 359. https://doi.org/10.1176/appi.ps.201100518Ruiter.

Coutanche, M. N., & Hallion, L. S. (2020). Machine learning for clinical psychology and clinical neuroscience. In A. G. C. Wright &

M. N. Hallquist (Eds.), *The Cambridge handbook of research methods in clinical psychology. Cambridge handbooks in psychology* (pp. 467–482). Cambridge: Cambridge University Press.

Cox, G. R., Callahan, P., Churchill, R., Hunot, V., Merry, S. N., Parker, A. G., et al. (2014). Psychological therapies versus antidepressant medication, alone and in combination for depression in children and adolescents. *Cochrane Database of Systematic Reviews*. https://doi.org/10.1002/14651858.CD008324.pub3.

Crutzen, R., Ruiter, R. A. C., & de Vries, N. K. (2014). Can interest and enjoyment help to increase use of Internet-delivered interventions? *Psychology & Health, 29*(11), 1227–1244. https://doi.org/10.1080/08870446.2014.921300.

Cuijpers, P., Donker, T., van Straten, A., Li, J., & Andersson, G. (2010). Is guided self-help as effective as face-to-face psychotherapy for depression and anxiety disorders? A systematic review and meta-analysis of comparative outcome studies. *Psychological Medicine, 40*, 1943–1957. https://doi.org/10.1017/S0033291710000772.

Cuijpers, P., Noma, H., Karyotaki, E., Cipriani, A., & Furukawa, T. A. (2019). Effectiveness and acceptability of cognitive behavior therapy delivery formats in adults with depression: A network meta-analysis. *JAMA Psychiatry, 76*(7), 700–707. https://doi.org/10.1001/jamapsychiatry.2019.0268.

Dagnea, G. A., Hendricks Brown, D., Howe, G., Kellam, S. G., & Liu, L. (2016). Testing moderation in network meta-analysis with individual participant data. *Statistics in Medicine, 35*(15), 2485–2502. https://doi.org/10.1002/sim.6883.

D'Acunto, G., Nageye, F., Zhang, J., Masi, G., & Cortese, S. (2019). Inflammatory cytokines in children and adolescents with depressive disorders: A systematic review and meta-analysis. *Journal of Child and Adolescent Psychopharmacology, 29*(5), 362–369. https://doi.org/10.1089/cap.2019.0015.

D'Amour, A. (2019). On multi-cause causal inference with unobserved confounding: Counterexamples, impossibility, and alternatives. Cornell University. https://arxiv.org/abs/1902.10286

De Choudhury, M., & Kiciman, E. (2018). Integrating artificial and human intelligence in complex, sensitive problem domains: Experiences from mental health. *AI Magazine, 39*(3), 69–80. https://doi.org/10.1609/aimag.v39i3.2815.

De Los Reyes, A., & Ohannessian, C. M. (2016). Introduction to the special issue: Discrepancies in adolescent-parent perceptions of the family and adolescent adjustment. *Journal of Youth and Adolescence, 45*(10), 1957–1972.

Deaton, A., & Cartwright, N. (2018). Reflections on randomized control trials. *Social Science & Medicine, 210*, 86–90. https://doi.org/10.1016/j.socscimed.2018.04.046.

Desai, R. J., Wang, S. V., Vaduganathan, M., Evers, T., & Schneeweiss, S. (2020). Comparison of machine learning methods with traditional models for use of administrative claims with electronic medical records to predict heart failure outcomes. *JAMA Network Open, 3*(1), e1918962. https://doi.org/10.1001/jamanetworkopen.2019.18962.

Dhruva, S. S., Ross, J. S., & Desai, N. R. (2018). Real-world evidence: Promise and peril for medical product evaluation. *P & T: A Peer-Reviewed Journal for Formulary Management, 43*(8), 464–472.

Doraiswamy, P. M., Blease, C., & Bodner, K. (2020). Artificial intelligence and the future of psychiatry: Insights from a global physician survey. *Artificial Intelligence in Medicine, 102*, 101753. https://doi.org/10.1016/j.artmed.2019.101753.

Du, J., Zhang, Y., Luo, J., Jia, Y., Wei, Q., Tao, C., et al. (2018). Extracting psychiatric stressors for suicide from social media using deep learning. *BMC Medical Informatics and Decision Making*. https://doi.org/10.1186/s12911-018-0632-8.

Durstewitz, D., Koppe, G., & Meyer-Lindenberg, A. (2019). Deep neural networks in psychiatry. *Molecular Psychiatry*. https://doi.org/10.1038/s41380-019-0365-9.

Duwe, G., & Kim, K. (2017). Out with the old and in with the new? An empirical comparison of supervised learning algorithms to predict recidivism. *Criminal Justice Policy Review, 28*, 570–600. https://doi.org/10.1177/0887403415604899.

Dwyer, D. B., Falkai, P., & Koutsouleris, N. (2018). Machine learning approaches for clinical psychology and psychiatry. *Annual Review of Clinical Psychology, 14*(1), 91–118. https://doi.org/10.1146/annurev-clinpsy-032816-045037.

Dyason, K. M., Shanley, D. C., O'Donovan, A., & Low-Choy, S. (2020). Does feedback improve psychotherapy outcomes compared to treatment-as-usual for adults and youth? *Psychotherapy Research, 30*(3), 310–324. https://doi.org/10.1080/10503307.2019.1620367.

Dyer, K., Hooke, G. R., & Page, A. C. (2016). Effects of providing domain specific progress monitoring and feedback to therapists and patients on outcome. *Psychotherapy Research*. https://doi.org/10.1080/10503307.2014.983207.

Elliott, M. L., Knodt, A. R., Ireland, D., Morris, M. L., Poulton, R., Ramrakha, S., et al. (2020). What is the test-retest reliability of common task-functional MRI measures? New empirical evidence and a meta-analysis. *Psychological Science, 10*, 15. https://doi.org/10.1177/0956797620916786.

Elliott, R., Wagner, J., Sales, C. M. D., Rodgers, B., Alves, P., & Café, M. J. (2016). Psychometrics of the personal questionnaire: A client-generated outcome measure. *Psychological Assessment, 28*(3), 263–278. https://doi.org/10.1037/pas0000174.

Emanuel, E. J., & Wachter, R. M. (2019). Artificial intelligence in health care: Will the value match the hype? *JAMA*. https://doi.org/10.1001/jama.2019.4914.

Esponda, G. M., Hartman, S., Qureshi, O., Sadler, E., Cohen, A., & Kakuma, R. (2020). Barriers and facilitators of mental health programmes in primary care in low-income and middle-income countries. *Lancet Psychiatry, 7*, 78–92. https://doi.org/10.1016/S2215-0366(19)30125-7.

Esteva, A., Kuprel, B., Novoa, R. A., Ko, J., Swetter, S. M., Blau, H. M., et al. (2017). Dermatologist-level classification of skin cancer with deep neural networks. *Nature, 542*, 115–118. https://doi.org/10.1038/nature21056.

Eyre, H. A., Becker, E. R. B., Blumenthal, M. S., Singh, A. B., Raji, C., Vahabzadeh, A., et al. (2020). Consumer participation in personalized psychiatry. In B. T. Baun (Ed.), *Personalized psychiatry* (pp. 63–68). San Diego: Academic Press.

Fernandes, B. S., Williams, L. M., Steiner, J., Leboyer, M., Carvalho, A. F., & Berk, M. (2017). The new field of "precision psychiatry". *BMC Medicine, 15*(1), 80. https://doi.org/10.1186/s12916-017-0849-x.

Fisher, A. J., Bosley, H. G., Fernandez, K. C., Reeves, J. W., Soystera, P., Diamond, A. E., et al. (2019). Open trial of a personalized modular treatment for mood and anxiety. *Behaviour Research and Therapy, 116*, 69–79.

Floridi, L. (2020). AI and Its new winter: From myths to realities. *Philosophy & Technology, 33*, 1–3. https://doi.org/10.1007/s13347-020-00396-6.

Flückiger, C., Del Re, A. C., Wampold, B. E., & Horvath, A. O. (2018). The alliance in adult psychotherapy: A meta-analytic synthesis. *Psychotherapy, 55*(4), 316–340. https://doi.org/10.1037/pst0000172.

Fonagy, P., & Allison, E. (2017). Commentary: A refresh for evidence-based psychological therapies-reflections on Marchette and Weisz. *Journal of Child Psychology and Psychiatry and Allied Disciplines, 58*(9), 985–987. https://doi.org/10.1111/jcpp.12805.

Foster, E. M., & Bickman, L. (2000). Refining the costs analyses of the Fort Bragg Evaluation: The impact of cost offset and cost shifting. *Mental Health Services Research, 2*(1), 13–25. https://doi.org/10.1023/A:1010139823791.

Franklin, C., Kim, J. S., Beretvas, T. S., Zhang, A., Guz, S., Park, S., et al. (2017a). The effectiveness of psychosocial interventions delivered by teachers in schools: A systematic review and meta-analysis. *Clinical Child and Family Psychology Review, 20*(3), 333–350. https://doi.org/10.1007/s10567-017-0235-4.

Franklin, J. C., Ribeiro, J. D., Fox, K. R., Bentley, K. H., Kleiman, E. M., Huang, X., et al. (2017b). Risk factors for suicidal thoughts and behaviors: A meta-analysis of 50 years of research. *Psychological Bulletin, 143*(2), 187–232. https://doi.org/10.1037/bul0000084.

Furman, D., Campisi, J., Verdin, E., Carrera-Bastos, P., Targ, S., Franceschi, C., et al. (2019). Chronic inflammation in the etiology of disease across the life span. *Nature Medicine, 25*(12), 1822–1832. https://doi.org/10.1038/s41591-019-0675-0.

Gaffney, H., Mansell, W., & Tai, S. (2019). Conversational agents in the treatment of mental health problems: Mixed-method systematic review. *JMIR Mental Health, 6*(10), e14166. https://doi.org/10.2196/14166.

Galatzer-Levy, I. R., Ma, S., Statnikov, A., Yehuda, R., & Shalev, A. Y. (2017). Utilization of machine learning for prediction of post-traumatic stress: A re-examination of cortisol in the prediction and pathways to non-remitting PTSD. *Translational Psychiatry, 7*, e1070. https://doi.org/10.1038/tp.2017.38.

Gambhir, S. S., Ge, T. J., Vermesh, O., & Spitler, R. (2018). Toward achieving precision health. *Science Translational Medicine*. https://doi.org/10.1126/scitranslmed.aao3612.

Garb, H. N., & Wood, J. M. (2019). Methodological advances in statistical prediction. *Psychological Assessment, 31*(12), 1456–1466. https://doi.org/10.1037/pas0000673.

Gargeya, R., & Leng, T. (2017). Automated identification of diabetic retinopathy using deep learning. *Ophthalmology*, *124*(7), 962–969. https://doi.org/10.1016/j.ophtha.2017.02.008.

Garland, A. F., Bickman, L., & Chorpita, B. F. (2010). Change what? Identifying quality improvement targets by investigating usual mental health care. *Administration and Policy in Mental Health and Mental Health Services Research, 37*(1–2), 15–26. https://doi.org/10.1007/s10488-010-0279-y.

Gharani, P., Suffoletto, B., Chung, T., & Karimi, H. A. (2017). An artificial neural network for movement pattern analysis to estimate blood alcohol content level. *Sensors*. https://doi.org/10.3390/s17122897.

Gómez Penedo, J. M., Babl, A. M., grosse Holtforth, M., Hohagen, F., Krieger, T., Lutz, W., et al. (2020). The association of therapeutic alliance with long-term outcome in a guided internet intervention for depression: Secondary analysis from a randomized control trial. *Journal of Medical Internet Research, 22*(3), e15824. https://doi.org/10.2196/15824.

Gondek, D., Edbrooke-Childs, J., Fink, E., Deighton, J., & Wolpert, M. (2016). Feedback from outcome measures and treatment effectiveness, treatment efficiency, and collaborative practice: A systematic review. *Administration and Policy in Mental Health and Mental Health Services Research, 43*(3), 325–343. https://doi.org/10.1007/s10488-015-0710-5.

Goodyear, R. K., Wampold, B. E., Tracey, T. J. G., & Lichtenberg, J. W. (2017). Psychotherapy expertise should mean superior outcomes and demonstrable improvement over time. *The Counseling Psychologist, 45*(1), 54–65. https://doi.org/10.1177/0011000016652691.

Gould, C. E., Kok, B. C., Ma, V. K., Zapata, A. M. L., Owen, J. E., & Kuhn, E. (2018). Veterans affairs and the department of defense mental health apps: A systematic literature review. *Psychological Services, 16*(2), 196–207. https://doi.org/10.1037/ser0000289.

Gual-Montolio, P., Martínez-Borba, V., Bretón-López, J. M., Osma, J., & Suso-Ribera, C. (2020). How are information and communication technologies supporting routine outcome monitoring and measurement-based care in psychotherapy? A systematic review. *International Journal of Environmental Research and Public Health, 17*, 3170. https://doi.org/10.3390/ijerph17093170.

Guntuku, S. C., Yaden, D. B., Kern, M. L., Ungar, L. H., & Eichstaedt, J. C. (2017). Detecting depression and mental illness on social media: An integrative review. *Current Opinion in Behavioral Sciences, 18*, 43–49. https://doi.org/10.1016/j.cobeha.2017.07.005.

Gyani, A., Shafran, R., Myles, P., & Rose, S. (2014). The gap between science and practice: How therapists make their clinical decisions. *Behavior Therapy, 45*(2), 199–211. https://doi.org/10.1016/j.beth.2013.10.004.

Hagendorff, T., & Wezel, K. (2019). 15 challenges for AI: Or what AI (currently) can't do. *AI & Society*. https://doi.org/10.1007/s00146-019-00886-y.

Hall, N. S. (2007). R. A. Fisher and his advocacy of randomization. *Journal of the History of Biology, 40*, 295–325. https://doi.org/10.1007/s10739-006-9119-z.

Harari, G. M., Müller, S. R., Aung, M. S., & Rentfrow, P. J. (2017). Smartphone sensing methods for studying behavior in everyday life. *Current Opinion in Behavioral Sciences, 18*, 83–90. https://doi.org/10.1016/j.cobeha.2017.07.018.

Hariman, K., Ventriglio, A., & Bhurga, D. (2019). The future of digital psychiatry. *Current Psychiatry Reports*. https://doi.org/10.1007/s11920-019-1074-4.

Hartley, W., Raphael, J., Lovell, K., & Berry, K. (2020). Effective nurse–patient relationships in mental health care: A systematic review of interventions to improve the therapeutic alliance. *International Journal of Nursing Studies, 102*, 103490. https://doi.org/10.1016/j.ijnurstu.2019.103490.

Hasanzadeh, F., Mohebbi, M., & Rostami, R. (2019). Prediction of rTMS treatment response in major depressive disorder using machine learning techniques and nonlinear features of EEG signal. *Journal of Affective Disorders, 256*, 132–142. https://doi.org/10.1016/j.jad.2019.05.070.

Hassani, H., Huang, X., & Ghodsi, M. (2018). Big data and causality. *Annals of Data Science, 5*, 133–156. https://doi.org/10.1007/s40745-017-0122-3.

Heckers, S. (2015). The value of psychiatric diagnoses. *JAMA Psychiatry, 72*(12), 1165–1166. https://doi.org/10.1001/jamapsychiatry.2015.2250.

Hedrick, T. E., Bickman, L., & Rog, D. J. (1993). *Applied research design: A practical guide*. Newbury Park, CA: Sage.

Heflinger, C. A., & Bickman, L. (1996). Family empowerment: A theoretically driven intervention and evaluation. In C. A. Heflinger & C. Nixon (Eds.), *Families and mental health services for children and adolescents* (pp. 96–116). Thousand Oaks: Sage.

Hermes, E. D. A., Lyon, A. R., Schueller, S. M., & Glass, J. E. (2019). Measuring the implementation of behavioral intervention technologies: Recharacterization of established outcomes. *Journal of Medical Internet Research, 21*(1), e11752. https://doi.org/10.2196/11752.

Hill, C. E., Spiegel, S. B., Hoffman, M. A., Kivlighan, D. M., & Gelso, C. J. (2017). Therapist expertise in psychotherapy revisited. *The Counseling Psychologist, 45*(1), 7–53. https://doi.org/10.1177/0011000016641192.

Hinton, G. (2018). Deep learning: A technology with the potential to transform health care. *JAMA, 320*(11), 1101–1102. https://doi.org/10.1001/jama.2018.11100.

Ho, C. S., Chee, C. Y., & Ho, R. C. (2020). Mental health strategies to combat the psychological impact of COVID-19 beyond paranoia and panic. *ANNALS: Academy of Medicine Singapore, 49*, 155–160.

Hoagwood, K. (1997). Interpreting nullity: The Fort Bragg experiment: A comparative success or failure? *American Psychologist, 52*(5), 546–550. https://doi.org/10.1037/0003-066X.52.5.546-550.

Hodgkinson, S., Godoy, L., Beers, L. S., & Lewin, A. (2017). Improving mental health access for low-income children and families in the primary care setting. *Pediatrics, 139*(1), 1–9. https://doi.org/10.1542/peds.2015-1175.

Holden, E. W., Friedman, R. M., & Santiago, R. L. (2001). Overview of the National Evaluation of the comprehensive community mental health services for children and their families program. *Journal of Emotional and Behavioral Disorders, 9*(1), 4–12. https://doi.org/10.1177/106342660100900102.

Holmes, E. A., O'Connor, R. C., Perry, V. H., Tracey, I., Wessely, S., Arseneault, L., et al. (2020). Multidisciplinary research priorities for the COVID-19 pandemic: A call for action for mental health science. *The Lancet Psychiatry, 7*(6), 547–560. https://doi.org/10.1016/S2215-0366(20)30168-1.

Hooke, G. R., Sng, A. A., Cunningham, N. K., & Page, A. C. (2017). Methods of delivering progress feedback to optimise patient outcomes: The value of expected treatment trajectories. *Cognitive Therapy and Research, 42*, 204–211.

Hopp, W. J., Li, J., & Wang, G. (2018). Big data and the precision medicine revolution. *Production and Operations Management, 27*(9), 1647–1664. https://doi.org/10.1111/poms.12891.

Hunt, G. E., Siegfried, N., Morley, K., Sitharthan, T., & Cleary, M. (2013). Psychosocial interventions for people with both severe mental illness and substance misuse. *Cochrane Database of Systematic Reviews*. https://doi.org/10.1002/14651858.CD001088.pub3.

Imel, Z. E., Caperton, D. D., Tanana, M., & Atkins, D. C. (2017). Technology-enhanced human interaction in psychotherapy. *Journal of Counseling Psychology, 64*(4), 385–393. https://doi.org/10.1037/cou0000213.

Iniesta, R., Malki, K., Maier, W., Rietschel, M., Mors, O., Hauser, J., et al. (2016). Combining clinical variables to optimize prediction of antidepressant treatment outcomes. *Journal of Psychiatric Research, 78*, 94–102. https://doi.org/10.1016/j.jpsychires.2016.03.016.

Institute of Medicine. (2007). *IOM Roundtable on evidence-based medicine: The learning healthcare system: Workshop summary*. Washington, DC: National Academies Press.

Ionita, G., Fitzpatrick, M., Tomaro, J., Chen, V. V., & Overington, L. (2016). Challenges of using progress monitoring measures: Insights from practicing clinicians. *Journal of Counseling Psychology, 63*(2), 173–182. https://doi.org/10.1037/cou0000122.

Jacobucci, R., & Grimm, K. J. (2020). Machine learning and psychological research: The unexplored effect of measurement. *Perspectives on Psychological Science, 15*(3), 809–816. https://doi.org/10.1177/1745691620902467.

Jones, P. J., Mair, P., Kuppens, S., & Weisz, J. R. (2019). An upper limit to youth psychotherapy benefit? A meta-analytic copula approach to psychotherapy outcomes. *Clinical Psychological Science, 7*(6), 1434–1449. https://doi.org/10.1177/2167702619858424.

Kaiser, J. (2015, January 30). Obama gives East Room rollout to precision medicine initiative. *Science*.https://www.sciencemag.org/news/2015/01/obama-gives-east-room-rollout-precision-medicine-initiative

Kar, S. K. (2019). Predictors of response to repetitive transcranial magnetic stimulation in depression: A review of recent updates. *Clinical Psychopharmacology and Neuroscience, 17*(1), 25–33.

Kasthurirathne, S. N., Biondich, P. G., Grannis, S. J., Purkayastha, S., Vest, J. R., & Jones, J. F. (2019). Identification of patients in need of advanced care for depression using data extracted from a statewide health information exchange: A machine learning approach. *Journal of Medical Internet Research, 21*(7), e13809. https://doi.org/10.2196/13809.

Kaur, P., & Sharma, M. (2019). Diagnosis of human psychological disorders using supervised learning and nature-inspired computing techniques: A meta-analysis. *Journal of Medical Systems*. https://doi.org/10.1007/s10916-019-1341-2.

Kazdin, A. E. (2019). Annual research review: Expanding mental health services through novel models of intervention delivery. *Journal of Child Psychology and Psychiatry and Allied Disciplines, 60*(4), 455–472. https://doi.org/10.1111/jcpp.12937.

Kee, F., & Taylor-Robinson, D. J. (2020). *Journal of Epidemiology and Community Health*. https://doi.org/10.1136/jech-2019-213311.

Kelley, S. D., Vides de Andrade, A. R., Sheffer, E., & Bickman, L. (2010). Exploring the black box: Measuring youth treatment process and progress in usual care. *Administration and Policy in Mental Health and Mental Health Services Research, 37*(3), 287–300. https://doi.org/10.1007/s10488-010-0298-8.

Kelley, S. D., Vides de Andrade, A. R., Bickman, L., & Robin, A. (2012). The Session Report Form (SRF): Are clinicians addressing issues of concern to youth and caregivers? *Administration and Policy in Mental Health and Mental Health Services Research, 39*(1–2), 133–145. https://doi.org/10.1007/s10488-012-0415-y.

Kent, D. M., Steyerberg, E., & van Klaveren, D. (2018). Personalized evidence based medicine: Predictive approaches to heterogeneous treatment effects. *BMJ, 363*, k4245. https://doi.org/10.1136/bmj.k4245.

Kessler, R. C., Bossarte, R. M., Luedtke, A., Zaslavsky, A. M., & Zubizarreta, J. R. (2019a). Machine learning methods for developing precision treatment rules with observational data. *Behaviour Research and Therapy*. https://doi.org/10.1016/j.brat.2019.103412.

Kessler, R. C., Chalker, S. A., Luedtke, A. R., Sadikova, E., & Jobes, D. A. (2019b). A preliminary precision treatment rule for remission of suicide ideation. *Suicide and Life-Threatening Behavior, 50*(2), 558–572. https://doi.org/10.1111/sltb.12609.

Kessler, R. C., Warner, C. H., Ivany, C., Petukhova, M. V., Rose, S., Bromet, E. J., et al. (2015). Predicting suicides after psychiatric hospitalization in U.S. Army soldiers: The Army Study to Assess Risk and Resilience in Servicemembers (Army STARRS). *Journal of the American Medical Association Psychiatry, 72*, 49–57. https://doi.org/10.1001/jamapsychiatry.2014.1754.

King, R. (2017). [Special issue]. Therapist effects in mental health service outcome. *Administration and Policy in Mental Health and Mental Health Services Research, 44*(5), 595–816.

King, R., & Bickman, L. (2017). Is there a future for therapists? *Administration and Policy in Mental Health and Mental Health Services Research, 44*(5), 595–597. https://doi.org/10.1007/s10488-017-0814-1.

Kissinger, H. A., Schmidt, E., &, Huttenlocher, D. (2019). The metamorphosis. *The Atlantic.*

Kiyotani, K., Chan, H. T., & Nakamura, Y. (2018). Immunopharmacogenomics towards personalized cancer immunotherapy targeting neoantigens. *Cancer Science, 109*(3), 542–549. https://doi.org/10.1111/cas.13498.

Kleiman, E. M., Turner, B. J., Fedor, S., Beale, E. E., Huffman, J. C., & Nock, M. K. (2017). Examination of real-time fluctuations in suicidal ideation and its risk factors: Results from two ecological momentary assessment studies. *Journal of Abnormal Psychology, 126*(6), 726–738. https://doi.org/10.1037/abn0000273.

Kluger, A. N., & DeNisi, A. (1996). The effects of feedback interventions on performance: A historical review, a meta-analysis, and a preliminary feedback intervention theory. *Psychological Bulletin, 119*(2), 254–284. https://doi.org/10.1037/0033-2909.119.2.254.

Kohler, S. (2018). Precision medicine: Moving away from one-size-fits-all. *Quest: Science for South Africa, 14*(3). Academy of

Science of South Africa (ASSAf)). Available at https://hdl.handle.net/20.500.11911/103

Konig, M. F., Powell, M., Staedtke, V., Bai, R.-Y., Thomas, D. L., Fischer, N., et al. (2020). Preventing cytokine storm syndrome in COVID-19 using α-1 adrenergic receptor antagonists. *Journal of Clinical Investigation*. https://doi.org/10.1172/JCI139642.

Koutsouleris, N., Kambeitz-Ilankovic, L., Ruhrmann, S., Rosen, M., Ruef, A., Dwyer, D. B., et al. (2018a). Prediction models of functional outcomes for individuals in the clinical high-risk state for psychosis or with recent-onset depression: A multimodal, multisite machine learning analysis. *JAMA Psychiatry, 75*(11), 1156–1172. https://doi.org/10.1001/jamapsychiatry.2018.2165.

Koutsouleris, N., Wobrock, T., Guse, B., Langguth, B., Landgrebe, M., Eichhammer, P., et al. (2018b). Predicting response to repetitive transcranial magnetic stimulation in patients with schizophrenia using structural magnetic resonance imaging: A multisite machine learning analysis. *Schizophrenia Bulletin, 44*(5), 1021–1034. https://doi.org/10.1093/schbul/sbx114.

Krittanawong, C., Johnson, K. W., & Tang, W. W. (2019). How artificial intelligence could redefine clinical trials in cardiovascular medicine: Lessons learned from oncology. *Personalized Medicine, 16*(2), 87–92. https://doi.org/10.2217/pme-2018-0130.

Kuang, K., Li, L., Geng, Z., Xu, L., Zhang, K., Liao, B., et al. (2020). Casual inference. *Engineering, 6*(3), 253–263. https://doi.org/10.1016/j.eng.2019.08.016.

Lake, B. M., Ullman, T. D., Tenenbaum, J. B., & Gershman, S. J. (2016). Building machines that learn and think like people. *Behavioral and Brain Sciences, 40*, 1–58.

Lambert, M. J., Whipple, J. L., & Kleinstäuber, M. (2018). Collecting and delivering progress feedback: A meta-analysis of routine outcome monitoring. *Psychotherapy, 55*(4), 520–537. https://doi.org/10.1037/pst0000167.

Laraway, S., Snycerski, S., Pradhan, S., & Huitema, B. E. (2019). An overview of scientific reproducibility: Consideration of relevant issues for behavior science/analysis. *Perspectives on Behavior Science, 42*(1), 33–57. https://doi.org/10.1007/s40614-019-00193-3.

La Rosa, J. (2018). The $10 billion self-improvement market adjusts to a new generation. Market Research Blog. https://blog.marketresearch.com/the-10-billion-self-improvement-market-adjusts-to-new-generation

Lattie, E. G., Nicholas, J., Knapp, A. A., Skerl, J. J., Kaiser, S. M., & Mohr, D. C. (2019). Opportunities for and tensions surrounding the use of technology-enabled mental health services in community mental health care. *Administration and Policy in Mental Health and Mental Health Services Research*. https://doi.org/10.1007/s10488-019-00979-2.

Lawrie, S. M., Fletcher-Watson, S., Whalley, H. C., & McIntosh, A. M. (2019). Predicting major mental illness: Ethical and practical considerations. *BJPsych Open, 5*(2), 1–5. https://doi.org/10.1192/bjo.2019.11.

Lechner, M. (2018). Modified causal forests for estimating heterogeneous causal effects. Retrieved from https://arxiv.org/abs/1812.09487.

Lee, Y., Ragguett, R. M., Mansur, R. B., Boutilier, J. J., Rosenblat, J. D., Trevizol, A., et al. (2018). Applications of machine learning algorithms to predict therapeutic outcomes in depression: A meta-analysis and systematic review. *Journal of Affective Disorders, 241*, 519–532. https://doi.org/10.1016/j.jad.2018.08.073.

Leighton, S. P., Upthegrove, R., Krishnadas, R., Benros, M. E., Broome, M. R., Gkoutos, G. V., et al. (2019). Development and validation of multivariable prediction models of remission, recovery, and quality of life outcomes in people with first episode psychosis: a machine learning approach. *Lancet Digital Health, 1*(6), e261–e270.

Lenze, E. J., Rodebaugh, T. L., & Nicol, G. E. (2020). A framework for advancing precision medicine in clinical trials for mental disorders. *JAMA Psychiatry*. https://doi.org/10.1001/jamapsychiatry.2020.0114.

Lewis, C. C., Boyd, M., Puspitasari, A., Navarro, E., Howard, J., Kassab, H., et al. (2019). Implementing measurement-based care in behavioral health: A review. *JAMA Psychiatry, 76*(3), 324–335. https://doi.org/10.1001/jamapsychiatry.2018.3329.

Librenza-Garcia, D. (2019). Ethics in the era of big data. In *Personalized Psychiatry: Big Data Analytics in Mental Health* (pp. 161–172). https://doi.org/10.1007/978-3-030-03553-2_9

Liu, X., Faes, L., Kale, A. U., Wagner, S. K., Fu, D. J., Bruynseels, A., et al. (2019a). A comparison of deep learning performance against health-care professionals in detecting diseases from medical imaging: A systematic review and meta-analysis. *Lancet Digital Health*. https://doi.org/10.1016/S2589-7500(19)30123-2.

Liu, X., Luo, X., Jiang, C., & Zhao, H. (2019b). Difficulties and challenges in the development of precision medicine. *Clinical Genetics., 95*(5), 569–574. https://doi.org/10.1111/cge.13511.

Liu, Y., Chen, P.-H. C., Krause, J., & Peng, L. (2019). How to read articles that use machine learning: Users' guides to the medical literature. *JAMA, 322*(18), 1806–1816. https://doi.org/10.1001/jama.2019.16489.

Liu, Y., Gadepalli, K., Norouzi, M., Dahl, G. E., Kohlberger, T., Boyko, A., Venugopalan, S., Timofeev, A., Nelson, P. Q., Corrado, G. W., Hipp, J. D., Peng, L., & Stumpe, M. C. (2017). Detecting cancer metastases on gigapixel pathology images. Retrieved from https://arxiv.org/abs/1703.02442 [cs.CV]

Love-Koh, J., Peel, A., Rejon-Parrilla, J. C., Ennis, K., Lovett, R., Manca, A., et al. (2018). The future of precision medicine: Potential impacts for health technology assessment. *PharmacoEconomics, 36*(12), 1439–1451. https://doi.org/10.1007/s40273-018-0686-6.

Low, D. M., Bentley, K. H., & Ghosh, S. S. (2020). Automated assessment of psychiatric disorders using speech: A systematic review. *Laryngoscope Investigative Otolaryngology, 5*, 96–116. https://doi.org/10.1002/lio2.354.

Lutz, W., Rubel, J. A., Schwartz, B., Schilling, V., & Deisenhofer, A. K. (2019). Towards integrating personalized feedback research into clinical practice: Development of the Trier Treatment Navigator (TTN). *Behaviour Research and Therapy, 120*, 103438. https://doi.org/10.1016/j.brat.2019.103438.

Mackrill, T., & Sørensen, K. M. (2019). Implementing routine outcome measurement in psychosocial interventions: A systematic review. *European Journal of Social Work*. https://doi.org/10.1080/13691457.2019.1602029.

Magrabi, F., Ammenwerth, E., Brender, J. B., De Keizer, N. F., Hyppönen, H., Nykänen, N., et al. (2019). Artificial intelligence in clinical decision support: Challenges for evaluating AI and practical implications. *IMIA Yearbook of Medical Informatics, 28*(1), 128–134. https://doi.org/10.1055/s-0039-1677903.

Mainsky, D., & Danks, D. (2018). Causal discovery algorithms: A practical guide. *Philosophy Compass, 13*(1), e12470. https://doi.org/10.1111/phc3.12470.

Mak, K. K., Lee, K., & Park, C. (2019). Applications of machine learning in addiction studies: A systematic review. *Psychiatry Research, 275*, 53–60. https://doi.org/10.1016/j.psychres.2019.03.001.

Marchette, L. K., & Weisz, J. R. (2017). Practitioner review: Empirical evolution of youth psychotherapy toward transdiagnostic approaches. *Journal of Child Psychology and Psychiatry and Allied Disciplines, 58*(9), 970–984. https://doi.org/10.1111/jcpp.12747.

Marcus, G. (2018). Deep learning: A critical appraisal. https://arxiv.org/abs/1801.00631.

Marshall, J. M., Dunstan, D. A., & Bartick, W. (2020). Clinical or gimmickal: The use and effectiveness of mobile mental health apps for treating anxiety and depression. *Australian & New Zealand Journal of Psychiatry, 54*(1), 20–28. https://doi.org/10.1177/0004867419876700.

Martel, M. M., Markon, K., & Smith, G. T. (2017). Research review: Multi-informant integration in child and adolescent psychopathology diagnosis. *Journal of Child Psychology and Psychiatry, and Allied Disciplines, 58*(2), 116–128. https://doi.org/10.1111/jcpp.12611.

Matheny, M., Israni, S. T., Ahmed, M., & Whicher, D. (Eds.). (2019). *Artificial intelligence in health care: The hope, the hype, the promise, the peril. NAM Special Publication*. Washington, DC: National Academy of Medicine.

Mayo, L. M., & Heilig, M. (2019). In the face of stress: Interpreting individual differences in stress-induced facial expressions. *Neurobiology of Stress, 10*, 100166. https://doi.org/10.1016/j.ynstr.2019.100166.

McGinnis, R. S., McGinnis, E. W., Hruschak, J., Lopez-Duran, N. L., Fitzgerald, K., Rosenblum, K. L., et al. (2019). Rapid detection of internalizing diagnosis in young children enabled by wearable sensors and machine learning. *PLoS ONE, 14*(1), e0210267. https://doi.org/10.1371/journal.pone.0210267.

Mental Health America. (2018). The state of mental health in America. https://mhanational.org/issues/state-mental-health-america

Milne, D. N., McCabe, K. L., & Calvo, R. A. (2019). Improving moderator responsiveness in online peer support through automated triage. *Journal of Medical Internet Research, 21*(4), e11410. https://doi.org/10.2196/11410.

Minar, M. R., & Naher, J. (2018). Recent advances in deep learning: An overview. https://doi.org/10.13140/RG.2.2.24831.10403

Miner, A. S., Shah, N., Bullock, K. D., Arnow, B. A., Bailenson, J., & Hancock, J. (2019). Key considerations for incorporating conversational AI in psychotherapy. *Frontiers in Psychiatry, 10*, 746. https://doi.org/10.3389/fpsyt.2019.00746.

Mittelstadt, B. (2019). The ethics of biomedical 'big data' analytics. *Philosophy & Technology, 32*(1), 17–21. https://doi.org/10.1007/s13347-019-00344-z.

Mohr, D. C., Lattie, E. G., Tomasino, K. N., Kwasny, M. J., Kaiser, S. M., Gray, E. L., et al. (2019). A randomized noninferiority trial evaluating remotely-delivered stepped care for depression using internet cognitive behavioral therapy (CBT) and telephone CBT. *Behaviour Research and Therapy, 123*, 103485. https://doi.org/10.1016/j.brat.2019.103485.

Mohr, D. C., Zhang, M., & Schueller, S. M. (2017). Personal sensing: Understanding mental health using ubiquitous sensors and machine learning. *Annual Review of Clinical Psychology, 13*, 23–47. https://doi.org/10.1146/annurev-clinpsy-032816-044949.

Morales, L. J., Ma, S., Aliferis, C., & Saxe, G. N. (2018). 32.2 The complex etiology of PTSD in children with maltreatment. *Journal of the American Academy of Child & Adolescent Psychiatry, 57*(10), S319.

Moustafa, A. A., Diallo, T. M. O., Amoroso, N., Zaki, N., Hassan, M., & Alashwal, H. (2018). Applying big data methods to understanding human behavior and health. *Frontiers in Computational Neuroscience, 12*, 84. https://doi.org/10.3389/fncom.2018.00084.

Mutz, J., Carter, B., Hurlemann, R., Fu, C., & Young, A. H. (2019). Comparative efficacy and acceptability of non-surgical brain stimulation for the acute treatment of major depressive episodes in adults: Systematic review and network meta-analysis. *BMJ, 364*, 1079. https://doi.org/10.1136/bmj.l1079.

Nadin, M. (2017). Rethinking the experiment: Necessary (R)evolution. (2018). *AI & Society, 33*, 467–485. https://doi.org/10.1007/s00146-017-0705-8.

Nahum-Shani, I., & Almirall, D. (2019). An introduction to adaptive interventions and SMART designs in education (NCSER 2020-001). U.S. Department of Education. Washington, DC: National Center for Special Education Research. Retrieved June 15, 2020, from https://ies.ed.gov/ncser/pubs/

Nakazawa, D. J. (2020). The angel and the assassin: The tiny brain cell that changed the course of medicine. Ballantine.

National Institutes of Health, Central Resource for Grants and Funding Information. (2001). NIH policy and guidelines on the inclusion of women and minorities as subjects in clinical research. https://grants.nih.gov/policy/inclusion/women-and-minorities/guidelines.htm

Ng, M. Y., & Weisz, J. R. (2016). Annual research review: Building a science of personalized intervention for youth mental health. *Journal of Child Psychology and Psychiatry and Allied Disciplines, 57*(3), 216–236. https://doi.org/10.1111/jcpp.12470.

Nield, T. (2019). Is deep learning already hitting its limitations? And is another AI winter coming? Towards Data Science. https://towardsdatascience.com/is-deep-learning-already-hitting-its-limitations-c81826082ac3.

Noda, Y., Silverstein, W., Barr, M., Vila-Rodriguez, F., Downar, J., Rajji, T., et al. (2015). Neurobiological mechanisms of repetitive transcranial magnetic stimulation of the dorsolateral prefrontal cortex in depression: A systematic review. *Psychological Medicine, 45*(16), 3411–3432. https://doi.org/10.1017/S0033291715001609.

O'Leary, M., Krailo, M., Anderson, J. R., & Reaman, G. H. (2008). Progress in childhood cancer: 50 years of research collaboration, a report from the Children's Oncology Group. *Seminars in Oncology, 35*(5), 484–493. https://doi.org/10.1053/j.seminoncol.2008.07.008.

Obermeyer, Z., Powers, B., Vogeli, C., & Mullainathan, S. (2019). Dissecting racial bias in an algorithm used to manage the health of populations. *Science, 366*(6464), 447–453. https://doi.org/10.1126/science.aax2342.

Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science*. https://doi.org/10.1126/science.aac4716.

Packin, N. G., & Lev-Aretz, Y. (2018). Learning algorithms and discrimination. In W. Barfield & U. Pagallo (Eds.), *Research handbook on the law of artificial intelligence*. Cheltenham, UK: Northampton.

Pardo, A., Jovanovic, J., Dawson, S., Gašević, D., & Mirriahi, N. (2019). Using learning analytics to scale the provision of personalised feedback. *British Journal of Educational Technology, 50*(1), 128–138. https://doi.org/10.1111/bjet.12592.

Park, A. L., Chorpita, B. F., Regan, J., Weisz, J. R., & The Research Network on Youth Mental Health. (2015). Integrity of evidence-based practice: Are providers modifying practice content or practice sequencing? *Administration and Policy in Mental Health and Mental Health Services Research, 42*(2), 186–196. https://doi.org/10.1007/s10488-014-0559-z.

Paulus, M. P., & Thompson, W. K. (2019). Computational approaches and machine learning for individual-level treatment predictions. *Psychopharmacology (Berl)*. https://doi.org/10.1007/s00213-019-05282-4.

Peake, J. M., Kerr, G., & Sullivan, J. P. (2018). A critical review of consumer wearables, mobile applications, and equipment for providing biofeedback, monitoring stress, and sleep in physically active populations. *Frontiers in Physiology, 9*, 743.

Pearl, J. (2000). *Causality: Models, reasoning, and inference*. Cambridge: Cambridge University Press.

Pearl, J. (2009). Causal inference in statistics: An overview. *Statistics Survey, 3*, 96–146. https://doi.org/10.1214/09-SS057.

Pearl, J. (2019). The seven tools of causal inference, with reflections on machine learning. *Communications of the ACM, 62*(3), 54–60. https://doi.org/10.1145/3241036.

Pearson, R., Pisner, D., Meyer, B., Shumake, J., & Beevers, C. G. (2018). A machine learning ensemble to predict treatment outcomes following an Internet intervention for depression. *Psychological Medicine, 49*(14), 2330–2341. https://doi.org/10.1017/S003329171800315X.

Pelham, W. E., Petras, H., & Pardini, D. A. (2020). Can machine learning improve screening for targeted delinquency prevention programs? *Prevention Science, 21*, 158–170. https://doi.org/10.1007/s11121-019-01040-2.

Peng, Z., Zho, C., Xue, S., Bai, J., Yu, S., Li, X., et al. (2018). Mechanism of repetitive transcranial magnetic stimulation for depression. *Shanghai Archives of Psychiatry, 30*(2), 84–92. https://doi.org/10.11919/j.issn.1002-0829.217047.

Pérez-Rosas, V., Wu, X., Resnicow, K., & Mihalcea, R. (2019). What makes a good counselor? Learning to distinguish between high-quality and low-quality counseling conversations. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics* (pp. 926–935). https://doi.org/10.18653/v1/p19-1088

Perkins, A., Ridler, J., Browes, D., Peryer, G., Notley, C., & Hackmann, C. (2018). Experiencing mental health diagnosis: A systematic review of service user, clinician, and carer perspectives across clinical settings. *The Lancet Psychiatry, 5*(9), 747–764. https://doi.org/10.1016/S2215-0366(18)30095-6.

Perlis, R. H. (2016). Abandoning personalization to get to precision in the pharmacotherapy of depression. *World Psychiatry, 15*(3), 228–235. https://doi.org/10.1002/wps.20345.

Perrin, A. J., & Pariante, C. M. (2020). Endocrine and immune effects of non-convulsive neurostimulation in depression: A systematic review. *Brain, Behavior, and Immunity.*. https://doi.org/10.1016/j.bbi.2020.02.016.

Pigoni, A., Delvecchio, G., Madonna, D., Bressi, C., Soares, J., & Brambilla, P. (2019). Can machine learning help us in dealing with treatment resistant depression? A review. *Journal of Affective Disorders, 259*, 21–26. https://doi.org/10.1016/j.jad.2019.08.009.

Pistorius, C. (2017). Developments in emerging digital health technologies. *DeltaHedron Innovation Insight,* No 1.1/17. https://www.deltahedron.co.uk/wp-content/uploads/2017/04/DeltaHedron_Innovation-Insight_Digital-health_No-1.1-17_-April-2017.pdf

Price, W. N. (2019). Potential liability for physicians using artificial intelligence. *JAMA, 322*(18), 1765–1766. https://doi.org/10.1001/jama.2019.15064.

Qiu, J., Shen, B., Zhao, M., Wang, Z., Xie, B., & Xu, Y. (2020). A nationwide survey of psychological distress among Chinese people in the COVID-19 epidemic: Implications and policy recommendations. *General Psychiatry, 33*(2), e100213. https://doi.org/10.1136/gpsych-2020-100213.

Quiroz, J. C., Geangu, E., & Yong, M. H. (2018). Emotion recognition using smart watch sensor data: Mixed-design study. *Journal of Medical Internet Research*. https://doi.org/10.2196/10153.

Rajkumar, R. P. (2020). COVID-19 and mental health: A review of the existing literature. *Asian Journal of Psychiatry, 52*, 102066. https://doi.org/10.1016/j.ajp.2020.102066.

Rajpurkar, P., Irvin, J., Zhu, K., Yang, B., Mehta, H., Duan, T., Ding, D., Bagul, A., Langlotz, C., Shpanskaya, K., Lungren, M. P., & Ng, A. Y. (2017). CheXNet: Radiologist-level pneumonia detection on chest X-rays with deep learning. arXiv:1711.05225 [cs.CV]. Retrieved from https://arxiv.org/abs/1711.05225

Ramsey, J., Glymour, M., Sanchez-Romero, R., & Glymour, C. (2016). A million variables and more: The fast greedy equivalence search algorithm for learning high-dimensional graphical causal models, with an application to functional magnetic resonance images. *International Journal of Data Science and Analytics, 3*(2), 121–129. https://doi.org/10.1007/s41060-016-0032-z.

Rauber, A., Trasarti, R., & Giannotti, F. (2019). Transparency in algorithmic decision making. Ercim News, (116), 10–11. Retrieved from www.ercim.eu: https://ercim-news.ercim.eu/en116/special/transparency-in-algorithmic-decision-making-introduction-to-the-special-theme

Rezaii, N., Walker, E., & Wolff, P. (2019). A machine learning approach to predicting psychosis using semantic density and latent content analysis. *npj Schizophrenia, 5*(9), 15. https://doi.org/10.1038/s41537-019-0077-9.

Riemer, M., Athay, M. M., Bickman, L., Breda, C., Kelley, S. D., & Vides de Andrade, A. R. (2012). The Peabody Treatment Progress Battery: History and methods for developing a comprehensive measurement battery for youth mental health. *Administration and Policy in Mental Health and Mental Health Services Research, 39*(1–2), 3–12. https://doi.org/10.1007/s10488-012-0404-1.

Riemer, M., & Bickman, L. (2011). Using program theory to link social psychology and program evaluation. In M. M. Mark, S. I. Donaldson, & B. Campbell (Eds.), *Social Psychology and Program/Policy Evaluation*. New York: Guilford.

Rodriguez-Ruiz, A., Lång, K., Gubern-Merida, A., Broeders, M., Gennaro, G., Clauser, P., et al. (2019). Stand-alone artificial intelligence for breast cancer detection in mammography: Comparison with 101 radiologists. *Journal of the National Cancer Institute, 111*(9), 9160922. https://doi.org/10.1093/jnci/djy222.

Rosenfeld, A., Benrimoh, D., Armstrong, C., Mirchi, N., Langlois-Therrien, T., Rollins, C., Tanguay-Sela, M., Mehltretter, J., Fratila, R., Israel, S., Snook, E., Perlman, K., Kleinerman, A., Saab, B., Thoburn, M., Gabbay, C., & Yaniv-Rosenfeld. (2019). Big data analytics and AI in mental healthcare. https://arxiv.org/abs/1903.12071.

Rudin, C., & Carlson, D. (2019). The secrets of machine learning: Ten things you wish you had known earlier to be more effective at data analysis. https://arxiv.org/abs/1906.01998.

Rush, J. A., & Ibrahim, H. M. (2018). Speculations on the future of psychiatric diagnosis. *Journal of Nervous and Mental Disease, 206*(6), 481–487. https://doi.org/10.1097/NMD.0000000000000821.

Rutledge, R. B., Chekroud, A. M., & Huys, Q. J. (2019). Machine learning and big data in psychiatry: Toward clinical applications. *Current Opinion in Neurobiology, 55*, 152–159. https://doi.org/10.1016/j.conb.2019.02.006.

Ryan, P., Luz, S., Albert, P., Vogel, C., Normand, C., & Elwyn, G. (2019). Using artificial intelligence to assess clinicians' communication skills. *BMJ (Online)*. https://doi.org/10.1136/bmj.l161.

Sachs, K., Perez, O., Pe'er, D., Lauffenburger, D. A., & Nolan, G. P. (2005). Causal protein-signaling networks derived from multiparameter single-cell data. *Science, 308*, 523–529. https://doi.org/10.1126/science.1105809.

Sales, C., Goncalves, S., Fragoeiro, A., Noronha, S., & Elliott, R. (2007). Psychotherapists openness to routine naturalistic idiographic research. *Mental Health and Learning Disabilities Research and Practice, 4*(2), 25. https://doi.org/10.5920/mhldrp.2007.42145.

Sales, C. M. D., & Alves, P. C. G. (2012). Individualized patient-progress systems: Why we need to move towards a personalized evaluation of psychological treatments. *Canadian Psychology, 53*(2), 115–121. https://doi.org/10.1037/a0028053.

Sales, C. M. D., & Alves, P. C. G. (2016). Patient-centered assessment in psychotherapy: A review of individualized tools. *Clinical Psychology Science and Practice, 23*(3), 265–283. https://doi.org/10.1111/cpsp.12162.

Sales, C. M. D., Alves, P. C. G., Evans, C., & Elliott, R. (2014). The Individualised Patient-Progress System: A decade of international collaborative networking. *Counselling and Psychotherapy*

Research, 14(3), 181–191. https://doi.org/10.1080/14733145.2014.929417.

Samek, W., Wiegand, T., & Müller, K.-R. (2017). Explainable artificial intelligence: Understanding, visualizing and interpreting deep learning models. https://arxiv.org/abs/1708.08296.

Sanchez, A. L., Cornacchio, D., Poznanski, B., Golik, A. M., Chou, T., & Comer, J. S. (2018). The effectiveness of school-based mental health services for elementary-aged children: A meta-analysis. *Journal of the American Academy of Child and Adolescent Psychiatry, 57*(3), 153–165. https://doi.org/10.1016/j.jaac.2017.11.022.

Saxe, G. N. (2019). Editorial: In the causal labyrinth: finding the path from early trauma to neurodevelopment. *Journal of the American Academy of Child and Adolescent Psychiatry, 58*(2), 159–163. https://doi.org/10.1016/j.jaac.2018.09.442.

Saxe, G. N. (2020). Redefining disease using informatics. In T. Adam & C. Aliferis (Eds.), *Personalized and precision medicine informatics. A workflow-based view*. Berlin: Springer.

Saxe, G. N., Ma, S, Morales, L. J., Galatzer-Levy, I., Aliferis, C., & Marmar, C. (in press). Computational causal discovery for post-traumatic stress in police officers**.** *Translational Psychiatry.*

Saxe, G. N., Ma, S., Ren, J., & Aliferis, C. (2017). Machine learning methods to predict child posttraumatic stress: A proof of concept study. *BMC Psychiatry*. https://doi.org/10.1186/s12888-017-1384-1.

Saxe, G. N., Statnikov, A., Fenyo, D., Ren, J., Li, Z., Prasad, M., et al. (2016). A complex systems approach to causal discovery in psychiatry. *PLoS ONE, 11*, e0151174. https://doi.org/10.1371/journal.pone.0151174.

Schmeidler, G. R. (1952). Personal values and ESP scores. *Journal of Abnormal and Social Psychology, 47*, 757−761. https://doi.org/10.1037/h0054954

Schneeberger, T., Ehrhardt, S., Anglet, M. S., & Gebhard, P. (2019). Would you follow my instructions if I was not human? Examining obedience towards virtual agents. In *Proceedings of the 2019 8th International Conference on Affective Computing and Intelligent Interaction (ACII)*.

Schork, N. J. (2015). Time for one-person trials. *Nature, 520*, 609–611. https://doi.org/10.1038/520609a.

Schuchmann, S. (2019). Probability of an approaching AI winter. *Towards data science.* https://towardsdatascience.com/probability-of-an-approaching-ai-winter-c2d818fb338a.

Schueller, S. M., Aguilera, A., & Mohr, D. C. (2017). Ecological momentary interventions for depression and anxiety. *Depression and Anxiety, 34*(6), 540–545. https://doi.org/10.1002/da.22649.

Sechrest, L., & Walsh, M. (1997). Dogma or data: Bragging rights. *American Psychologist, 52*(5), 536–540. https://doi.org/10.1037/0003-066X.52.5.536.

Serretti, A. (2018). The present and future of precision medicine in psychiatry: Focus on clinical psychopharmacology of antidepressants. *Clinical Psychopharmacology and Neuroscience, 16*(1), 1–6. https://doi.org/10.9758/cpn.2018.16.1.1.

Shatte, A. B. R., Hutchinson, D. M., & Teague, S. J. (2019). Machine learning in mental health: A scoping review of methods and applications. *Psychological Medicine, 49*(9), 1426–1448. https://doi.org/10.1017/S0033291719000151.

Shields, G. S., Spahr, C. M., & Slavich, G. M. (2020). Psychosocial interventions and immune system function: A systematic review and meta-analysis of randomized clinical trials. *JAMA Psychiatry*. https://doi.org/10.1001/jamapsychiatry.2020.0431.

Shimizu, S. (2019). Non-Gaussian methods for causal structure learning. *Prevention Science, 20*, 431–441. https://doi.org/10.1007/s11121-018-0901-x.

Shore, J. H., Schneck, C. D., & Mishkind, M. C. (2020). Telepsychiatry and the coronavirus disease 2019 pandemic—Current and future outcomes of the rapid virtualization of psychiatric care. *JAMA Psychiatry*. https://doi.org/10.1001/jamapsychiatry.2020.1643.

Shrivastava, A., Tripathy, A. K., & Dalal, P. K. (2019). A SVM-based classification approach for obsessive compulsive disorder by oxidative stress biomarkers. *Journal of Computational Science, 36*, 101023. https://doi.org/10.1016/j.jocs.2019.07.010.

Simon, G. E., & Yarborough, B. J. (2020). Good news: Artificial intelligence in psychiatry is actually neither. *Psychiatric Services, 71*(3), 219–220. https://doi.org/10.1176/appi.ps.201900464.

Slavich, G. M. (2019). Psychoneuroimmunology of stress and mental health. In K. L. Harkness & E. P. Hayden (Eds.), *The Oxford handbook of stress and mental health*. Oxford: Oxford University Press.

Slavich, G. M. (2020). Social safety theory: A biologically based evolutionary perspective on life stress, health, and behavior. *Annual Review of Clinical Psychology, 16*(1), 265.

Slavich, G. M., & Irwin, M. R. (2014). From stress to inflammation and major depressive disorder: A social signal transduction theory of depression. *Psychological Bulletin, 140*, 774–815. https://doi.org/10.1037/a0035302

Somani, A., & Kar, S. K. (2019). Efficacy of repetitive transcranial magnetic stimulation in treatment-resistant depression: The evidence thus far. *General Psychiatry*. https://doi.org/10.1136/gpsych-2019-100074.

Sonsin-Diaz, N., Gottesman, R. F., Fracica, E., Walston, J., Windham, G., Knopman, D. S., et al. (2020). Chronic systemic inflammation is associated with symptoms of late-life depression: The ARIC study. *The American Journal of Geriatric Psychiatry, 28*(1), 87–98. https://doi.org/10.1016/j.jagp.2019.05.011.

Stambaugh, F. L., Mustillo, S. A., Burns, B. J., Stephens, R. L., Baxter, B., Edwards, D., et al. (2007). Outcomes from wraparound and multisystemic therapy in a center for mental health services system-of-care demonstration site. *Journal of Emotional and Behavioral Disorders, 15*(3), 143–155. https://doi.org/10.1177/10426070150030201.

Statnikov, A., Tsamardinos, I., Dosbayev, Y., & Aliferis, C. F. (2005). GEMS [Gene expression model selector]: A system for automated cancer diagnosis and biomarker discovery from microarray gene expression data. *International Journal of Medical Informatics, 74*(7–8), 491–503. https://doi.org/10.1016/j.ijmedinf.2005.05.002.

Steering Committee of the Physicians' Health Study Research Group. (1989). Final report on the aspirin component of the ongoing Physicians' Health Study. *New England Journal of Medicine, 321*, 129–135. https://doi.org/10.1056/NEJM198907203210301.

Stephens-Davidowitz, S. (2017). *Everybody lies: Big data, new data and what the Internet can tell us about who we really are.* HarperCollins.

Stockman, F., & Eligon, J. (2020). Cities ask if it's time to defund police and 'reimagine' public safety. *New York Times.*https://www.nytimes.com/2020/06/05/us/defund-police-floyd-protests.html

Stroul, B. A., & Friedman, R. (1986). *A system of care for children and youth with severe emotional disturbances* (Rev ed.). Washington, DC: CASSP Technical Assistance Center, Georgetown University Child Development Center.

Study finds psychiatric diagnosis to be 'scientifically meaningless.' (2019). *Neuroscience News*.

Subramanian, S. V., Kim, R., & Christakis, N. A. (2018). The "average" treatment effect: A construct ripe for retirement. A commentary on Deaton and Cartwright. *Social Science and Medicine, 210*, 77–82. https://doi.org/10.1016/j.socscimed.2018.04.027.

Substance Abuse and Mental Health Services Administration (SAMHSA). (2019). System of Care (SOC) expansion and sustainability grants. https://www.samhsa.gov/grants/grant-announcements/sm-19-009

Sullivan, G. M., & Feinn, R. (2012). Using effect size—or why the P value is not enough. *Journal of Graduate Medical Education, 4*(3), 279–282. https://doi.org/10.4300/JGME-D-12-00156.1.

Sundaravadivel, P., Kougianos, E., Mohanty, S. P., & Ganapathiraju, M. K. (2018). Everything you wanted to know about smart health care: Evaluating the different technologies and components of the Internet of Things for better health. *IEEE Consumer Electronics Magazine, 7*(1), 18–28. https://doi.org/10.1109/MCE.2017.2755378.

Taitsman, J. K., VanLandingham, A., & Grimm, C. A. (2020). Commercial influences on electronic health records and adverse effects on clinical decision making. *JAMA Internal Medicine*. https://doi.org/10.1001/jamainternmed.2020.1318.

Tan, J., Rollins, C. P. E., Israel, S., & Benrimoh, D. (2019). Primed for psychiatry: The role of artificial intelligence and machine learning in the optimization of depression treatment. *University of Toronto Medical Journal, 96*(1), 43–47.

Tanana, M., Hallgren, K. A., Imel, Z. E., Atkins, D. C., & Srikumar, V. (2016). A comparison of natural language processing methods for automated coding of motivational interviewing. *Journal of Substance Abuse Treatment, 65*, 43–50. https://doi.org/10.1016/j.jsat.2016.01.006.

Tanana, M. J., Soma, C. S., Srikumar, V., Atkins, D. C., & Imel, Z. E. (2019). Development and evaluation of ClientBot: Patient-like conversational agent to train basic counseling skills. *Journal of Medical Internet Research, 21*(7), e12529. https://doi.org/10.2196/12529.

Tandon, N., & Tandon, R. (2018). Will machine learning enable us to finally cut the Gordian knot of schizophrenia? *Schizophrenia Bulletin, 44*(5), 939–941. https://doi.org/10.1093/schbul/sby101.

Tanner-Smith, E. E., Durlak, J. A., & Marx, R. A. (2018). Empirically based mean effect size distributions for universal prevention programs targeting school-aged youth: A review of meta-analyses. *Prevention Science, 19*(8), 1091–1101. https://doi.org/10.1007/s11121-018-0942-1.

Tasca, G. A., Angus, L., Bonli, R., Drapeau, M., Fitzpatrick, M., Hunsley, J., et al. (2019). Outcome and progress monitoring in psychotherapy: Report of a Canadian Psychological Association Task Force. *Canadian Psychology, 60*(3), 165–177. https://doi.org/10.1037/cap0000181.

Thompson A., & Bodoni, S. (2020). Google CEO Thinks AI will be more profound change than fire. *Bloomberg News*. Retrieved June 20, 2020, from https://www.bloomberg.com/news/articles/2020-01-22/google-ceo-thinks-ai-is-more-profound-than-fire#:~:text=%E2%80%9CAI%2520is%2520one%2520of%2520the,in%2520Davos%2C%2520Switzerland%2520on%2520Wednesday.&text=%E2%80%9CAI%2520is%2520no%2520different%2520from%2520the%2520climate%2C%E2%80%9D%2520Pichai%2520said

Tiana, L., Si-Si, S., Long-Biao, C., Shi-Quan, W., Zheng-Wu, P., Qing-Rong, T., et al. (2020). Repetitive transcranial magnetic stimulation elicits antidepressant- and anxiolytic-like effect *via* nuclear factor-E2-related factor 2-mediated anti-inflammation mechanism in rats. *Neuroscience, 429*(1), 119–133. https://doi.org/10.1016/j.neuroscience.2019.12.025.

Topol, E. J. (2019a). *Deep medicine: How artificial intelligence can make healthcare human again*. New York: Basic Books.

Topol, E. J. (2019b). High-performance medicine: The convergence of human and artificial intelligence. *Nature Medicine, 25*, 44–56. https://doi.org/10.1038/s41591-018-0300-7.

Triantafyllidis, A. K., & Tsanas, A. (2019). Applications of machine learning in real-life digital health interventions: Review of the literature. *Journal of Medical Internet Research, 21*(4), 1–9. https://doi.org/10.2196/12286.

Tuerk, P. W., Schaeffer, C. M., McGuire, J. F., Larsen, M. A., Capobianco, N., & Piacentini, J. (2019). Adapting evidence-based treatments for digital technologies: A critical review of functions, tools, and the use of branded solutions. *Current Psychiatry Reports*. https://doi.org/10.1007/s11920-019-1092-2.

An understanding of AI's limitations is starting to sink in. (2020). *Technology Quarterly: Artificial Intelligence and Its Limits*, *The Economist.*

Vaidyam, A. N., Wisniewski, H., Halamka, J. D., Matcheri, S., Kashavan, M. S., & Torous, J. B. (2019). Chatbots and conversational agents in mental health: A review of the psychiatric landscape. *Canadian Journal of Psychiatry, 64*(7), 456–464. https://doi.org/10.1177/070674371982897710.1177/.

Valles-Colomer, M., Falony, G., Darzi, Y., Tigchelaar, E. F., Wang, J., Tito, R. Y., et al. (2019). The neuroactive potential of the human gut microbiota in quality of life and depression. *Nature Microbiology, 4*, 623–632. https://doi.org/10.1038/s41564-018-0337-x.

Van Daele, T., Karekla, M., Kassianos, A. P., Compare, A., Haddouk, L., Salgado, J., et al. (2020). Recommendations for policy and practice of telepsychotherapy and E-mental health in Europe and beyond. *Journal of Psychotherapy Integration, 30*(2), 160–173. https://doi.org/10.1037/int0000218.

Veale, M., & Binns, R. (2017). Fairer machine learning in the real world: Mitigating discrimination without collecting sensitive data. *Big Data & Society, 4*, 1–17. https://doi.org/10.1177/2053951717743530

Vigerland, S., Lenhard, F., Bonnert, M., Lalouni, M., Hedman, E., Ahlen, J., et al. (2016). Internet-delivered cognitive behavior therapy for children and adolescents: A systematic review and meta-analysis. *Clinical Psychology Review, 50*, 1–10. https://doi.org/10.1016/j.cpr.2016.09.005.

Vinge, V. (1993). How to survive in the post-human era. In *Interdisciplinary Science and Engineering in the Era of Cyberspace, Proceedings of the VISION-21 Symposium.* NASA Conference Proceeding 10129. https://ntrs.nasa.gov/archive/nasa/casi.ntrs.nasa.gov/19940022855.pdf

Voigt, J., Carpenter, L., & Leuchter, A. (2019). A systematic literature review of the clinical efficacy of repetitive transcranial magnetic stimulation (rTMS) in non-treatment resistant patients with major depressive disorder. *BMC Psychiatry, 19*(1), 13. https://doi.org/10.1186/s12888-018-1989-z.

Walch, K. (2019). Are we heading for another AI winter soon? *Forbes.*https://www.forbes.com/sites/cognitiveworld/2019/10/20/are-we-heading-for-another-ai-winter-soon/#783bf81256d6.

Wampold, B. E., & Imel, Z. E. (2015). *The great psychotherapy debate: The evidence for what makes psychotherapy work* (2nd ed.). New York: Routledge. https://doi.org/10.4324/9780203582015

Waszczuk, M. A., Zimmerman, M., Ruggero, C., Li, K., MacNamara, A., Weinberg, A., et al. (2017). What do clinicians treat: Diagnoses or symptoms? The incremental validity of a symptom-based, dimensional characterization of emotional disorders in predicting medication prescription patterns. *Comprehensive Psychiatry, 79*, 80–88. https://doi.org/10.1016/j.comppsych.2017.04.004.

Weisz, J. R., Doss, A. J., & Hawley, K. M. (2006). Evidence-based youth psychotherapies versus usual clinical care: A meta-analysis of direct comparisons. *American Psychologist, 61*(7), 671–689. https://doi.org/10.1037/0003-066X.61.7.671.

Weisz, J. R., Han, S. S., & Valeri, S. M. (1997). More of what? Issues raised by the Fort Bragg study. *American Psychologist, 52*(5), 541–545. https://doi.org/10.1037/0003-066X.52.5.541.

Weisz, J. R., Kuppens, S., Eckshtain, D., Ugueto, A. M., Hawley, K. M., & Jensen-Doss, A. (2013). Performance of evidence-based youth psychotherapies compared with usual clinical care: A multilevel meta-analysis. *JAMA Psychiatry, 70*(7), 750–761. https://doi.org/10.1001/jamapsychiatry.2013.1176.

Weisz, J. R., Kuppens, S., Ng, M. Y., Eckshtain, D., Ugueto, A. M., Vaughn-Coaxum, R., et al. (2017). What five decades of research tells us about the effects of youth psychological therapy: A multilevel meta-analysis and implications for science and practice. *The American psychologist, 72*(2), 79–117. https://doi.org/10.1037/a0040360.

Weisz, J. R., Kuppens, S., Ng, M. Y., Vaughn-Coaxum, R. A., Ugueto, A. M., Eckshtain, D., et al. (2019). Are psychotherapies for young people growing stronger? Tracking trends over time for youth anxiety, depression, attention-deficit/hyperactivity disorder, and conduct problems. *Perspectives on Psychological Science, 14*(2), 216–237. https://doi.org/10.1177/1745691618805436.

Wen, C. K. F., Schneider, S., Stone, A. A., & Spruijt-Metz, D. (2017). Compliance with mobile ecological momentary assessment protocols in children and adolescents: A systematic review and meta-analysis. *Journal of Medical Internet Research*. https://doi.org/10.2196/jmir.6641.

Weng, S. F., Reps, J., Kai, J., Garibaldi, J. M., & Qureshi, N. (2017). Can machine-learning improve cardiovascular risk prediction using routine clinical data? *PLoS ONE*. https://doi.org/10.1371/journal.pone.0174944.

Wiedermann, W., Dong, N., & von Eye, A. (2019). Advances in statistical methods for causal inference in prevention science: Introduction to the Special Section. *Prevention Science, 20*(3), 390–393. https://doi.org/10.1007/s11121-019-0978-x.

Woodyard, C. (2019, April 22). Elon Musk vows fully self-driving Teslas this year and 'robotaxis' ready next year. USA Today. Retrieved from https://www.usatoday.com/story/money/cars/2019/04/22/tesla-says-its-fully-self-driving-car-tech-autonomous/3540926002/

Wolff, J. G. (2018). Solutions to problems with deep learning. https://arxiv.org/abs/1801.05457.

Wright, J. H., Mishkind, M., Yeager, C. M., Eells, T. D., & Chan, S. R. (2019). Computer-assisted cognitive-behavior therapy and mobile apps for depression and anxiety. *Current Psychiatry Reports*. https://doi.org/10.1007/s11920-019-1031-2.

Wrzeszczynski, K. O., Frank, M. O., Koyama, T., Rhrissorrakrai, K., Robine, N., Utro, F., et al. (2017). Comparing sequencing assays and human-machine analyses in actionable genomics for glioblastoma. *Neurology Genetics*. https://doi.org/10.1212/NXG.0000000000000164.

Wu, C. S., Luedtke, A. R., Sadikova, E., Tsai, H.-J., Liao, S.-C., Liu, C.-C., et al. (2020). Development and validation of a machine learning individualized treatment rule in first-episode schizophrenia. *JAMA Network Open, 3*(2), e1921660. https://doi.org/10.1001/jamanetworkopen.2019.21660.

Wykes, T. (2019). Racing towards a digital paradise or a digital hell? (2019). *Journal of Mental Health, 28*(1), 1–3. https://doi.org/10.1080/09638237.2019.1581360.

Yeager, C. M. & Benight, C. C. (2018). If we build it, will they come? Issues of engagement with digital health interventions for trauma recovery. *mHealth, 4,* 37. https://doi.org/10.21037/mhealth.2018.08.04

Zhou, X., Snoswell, C. L., Harding, L. E., Bambling, M., Edirippulige, S., Bai, X., et al. (2020). The role of telehealth in reducing the mental health burden from COVID-19. *Telemedicine and e-Health*, *26*(4), 377–379. https://doi.org/10.1089/tmj.2020.0068.

Zilcha-Mano, S. (2017). Is the alliance really therapeutic? Revisiting this question in light of recent methodological advances. *American Psychologist, 72*(4), 311–325. https://doi.org/10.1037/a0040435.

Zilcha-Mano, S. (2019). Major developments in methods addressing for whom psychotherapy may work and why. *Psychotherapy Research, 29*, 693−708. https://doi.org/10.1080/10503307.2018.1429691