Behaviour Research and Therapy 123 (2019) 103506

Contents lists available at ScienceDirect

Behaviour Research and Therapy

journal homepage: www.elsevier.com/locate/brat

The history, current status, and possible future of precision mental health

Robert J. DeRubeis

University of Pennsylvania, Department of Psychology, Stephen A. Levin Building, 425 S. University Avenue, Philadelphia, PA, 19104-6018, USA

ARTICLE INFO

Keywords:
Precision medicine
Personalized medicine
Moderators of treatment response
Multivariable models
Machine learning

ABSTRACT

In evidence-based mental health practice, decisions must often be made for which there is little or no empirical basis. A common example of this is when there are multiple empirically supported interventions for a person with a given diagnosis, where the aim is to recommend the treatment most likely to be effective for that person. Data obtained from randomized clinical trials allow for the identification of patient characteristics that could be used to match patients to treatments. Historically, researchers have focused on individual moderators, single variables that interact statistically with treatment type, but these have rarely proved powerful enough to inform treatment decisions. Recently, researchers have begun to explore ways in which the use of multivariable algorithms might improve clinical decision-making. Common pitfalls have been identified, including the use of methods that provide overoptimistic estimates of the gains that can be expected from the applications of an algorithm in a clinical setting. It is too early to tell if these efforts will pay off and, if so, how much their use can increase the efficiency and effectiveness of mental health systems. It behooves the field to continue to learn and develop the most powerful methods that can produce generalizable knowledge that will advance the aims of precision mental health.

It is now 52 years since Gordon Paul (1967) posed a question that has been paraphrased as "What works for whom?" in many hundreds of publications that center on the outcomes of psychotherapy. The phrase continues to be invoked frequently in the titles of papers as well as in an influential series of books about psychotherapy, and for good reason. The full quote, "*What* treatment, by *whom*, is most effective for *this* individual with *that* specific problem, and under *which* set of circumstances?" asks for more detail than has ever been addressed in the empirical literatures that reference it. Instead, researchers have used the spirit of the quote primarily to foster interest in a specific kind of question, "When there are two or more evidence-based treatments for a person in a given category, such as Major Depressive Disorder (MDD), how might we use person characteristics to guide the choice among them?" This version of the question starts with a diagnosis, which already narrows the choices to evidence-based treatments for patients with that diagnosis. Within a category, however, decisions about which treatment a patient should receive have been left largely to clinical judgment, patient preference, and to practical considerations such as availability and cost. The effort to tailor a treatment course specifically for an individual patient, beyond a consideration of diagnosis, has been termed "personalized medicine," a phrase that was coming into use beginning around 1940, according to Google nGram (Michel et al., 2011). Research that aims to provide an empirical base for such decisions would continue to be referred to with this term until early in this century, when "precision medicine" slowly began to replace it. Although precision medicine is often identified specifically with physical medicine, and even more narrowly with the use of genetic testing to guide treatment decisions, it is being adopted more widely to refer to evidence-based recommendations that are based on patient-specific features.

Early in this decade my colleagues and I began to wrestle with an example of how treatment recommendations might be "personalized," or made more precise, in a data-driven manner. We had completed a randomized comparison of two very different treatments for MDD – antidepressant medications (ADM) and cognitive behavioral therapy (CBT) – known to yield similar average effects on the symptoms of depression. Indeed, in our randomized clinical trial (RCT) of adult outpatients with relatively severe manifestations of MDD, the average amount of change on the primary outcome measure was nearly identical between the treatments, and both treatments outperformed a placebo condition (DeRubeis et al., 2005). One way to understand this pattern is to suppose that any given patient's likelihood of benefitting from one of the treatments is similar to his or her likelihood of benefitting from the other. In this scenario, predictions made irrespective of treatment ("prognostic" predictions) could be used to array patients along a continuum from those expected to do very poorly in either treatment to those for whom good outcomes are expected irrespective of treatment condition (DeRubeis, Gelfand, German, Fournier, &

E-mail address: derubeis@psych.upenn.edu.

https://doi.org/10.1016/j.brat.2019.103506
Received 9 April 2019; Received in revised form 14 October 2019; Accepted 25 October 2019
Available online 06 November 2019
0005-7967/

Forand, 2014). But the treatments we studied are presumed to work largely through mechanisms specific to each treatment, even if it has been difficult to demonstrate that this is so. Insofar as the mechanisms differ between two treatments, it ought to be possible to identify patients who are more likely to benefit from one than from the other, on the idea that some patients would benefit from the changes in the mechanisms produced by one treatment more than they would from the other (Cohen & DeRubeis, 2018). We became determined to locate, refine, or develop methods that could be used to make this kind of distinction in the context we were working in as well as in other contexts, such as one psychotherapy versus another, in which two treatments provide substantial benefit but appear to do so via different mechanisms.

We imagined a method that would use data from RCTs to build models linking individual patients' characteristics to their outcomes. The utility of such models could then be tested for their ability to improve outcomes, in the aggregate, when implemented and compared to random assignment or other means of treatment assignment.

Moderators

Historically, papers reporting on the differential prediction (also termed "moderation") of treatment effects obtained in an RCT have focused on a single variable or, in a few cases, on multiple variables but with each considered in isolation. In a typical study, a baseline variable is chosen in connection with a theory of change for one or more of the treatments. The verdict as to whether the variable has acted as a moderator is given by a significance test of the interaction between the variable and treatment assignment. Papers reporting on individual moderator effects in mental health treatment research have been common in the past two decades. For example, Kessler et al. (2017) identified 24 investigations of single-variable moderator effects in MDD, of which two thirds were published between 2006 and 2014. Discussion sections of these papers have tended to focus on the theoretical rather than the practical importance of the moderator findings, with at most a passing mention of the potential for their application to clinical decision-making (e.g., Fournier et al., 2009).

For many reasons, some best viewed in hindsight, findings of significant individual moderators were never likely to inform practice. First, if the discovery of an individual moderator was to have a chance of changing clinical practice, it would need to be powerful enough to identify, on its own, which treatment a given patient would most likely respond to, and yet subtle enough to have escaped the notice of mental health scientists and practitioners prior to the discovery of the moderator. Indeed, individual moderators rarely account for enough variance to turn up as significant in multiple research studies, in part because most RCTs are not powered to detect modest interaction effects. The failure of individual moderators to replicate can be discouraging until one realizes that to the extent that there exist multiple independent moderating effects in a given context, none of them can account for much of the relevant variance. This description fits the pattern of findings, set out in several papers, from Project MATCH Research Group, (1998), a well-known randomized comparison of three treatments for alcoholism. Between the primary publication from MATCH and several subsequent papers, many individual moderators were identified, but none of them evidenced effects strong enough to recommend their use as a guide to clinical practice.

In our own publications, using data from the DeRubeis et al. (2005) study cited above, we at first focused on individual moderator effects (Fournier et al., 2008, 2009; Leykin et al., 2007). In the discussion section of Fournier et al. (2009), commenting on the implications of our having identified three moderators of the effects of ADM versus CBT (marital status, employment status, and severity of recent negative life events), we observed that our findings "could [broadly] suggest meaningful recommendations for treatment providers." However, as was typical of others' interpretations of their single variable moderator findings, we did not spell out how this information could be used to inform treatment decisions. In part, this was because, as was true of the individual moderator papers cited in Kessler et al. 2017; see also Cohen & DeRubeis, 2018, we did not attempt to combine the information in a way that would take into account these moderators and others (personality disorder, history of ADM treatment) that we had identified in earlier publications of findings from the same data (Fournier et al., 2008; Leykin et al., 2007).

Numerous other papers had reported on moderators of response to treatments for depression, but no single biomarker, demographic, personality, or clinical variable appeared to account for enough of the differential effects of the treatments to support valid personalized treatment recommendations (Cohen & DeRubeis, 2018). Unless there is a dominant variable, a person's score on any single measure will provide, at most, a hint as to which treatment is likely to be better for that person. That same person's score on a different moderator variable might point to the opposite treatment, and so on for the other moderators. All of this seemed obvious, once we thought about it, so the question we set out to address became: "How can information from multiple variables best be combined to inform clinically valid and meaningful treatment recommendations?" Our question allowed not only for the possibility that some persons with depression are much more likely to benefit from CBT than from ADM, and vice versa, but that others would be roughly equally likely to respond to one as the other. Only an approach that combines information from a set of relatively weak independent moderators appeared capable of contributing to precision mental health in the way we envisioned it.

At the time there had been very few efforts in mental health that reflected similar ambitions, and each was unrelated to the others. In 1996, Barber and Muenz published findings from their analysis of data from the Treatment for Depression Collaborative Research Program (TDCRP; Elkin, Parloff, Hadley, & Autry, 1985) in which participants with depression were randomized to (among other conditions) Interpersonal Therapy (IPT) or Cognitive Therapy (CT). One set of analyses revealed significant treatment by patient-characteristic (moderator) effects for marital status, trait obsessiveness, and trait avoidance. Barber and Muenz showed how patients' characteristics could be combined into a "matching score," using weights from a regression model that included all three moderator variables simultaneously, as well as the main effects of treatment and of each of the moderator variables. The matching score for each patient was calculated by multiplying the respective coefficients for each effect in the model by the patient's values on the respective variables. If the sign of the resulting score was positive, it suggested that CT was the preferred treatment for the individual. A negative sign indicated that IPT was more likely to be effective for that patient. Barber and Muenz went on to compare the outcomes of the two treatments in each of two subsets. Those with positive scores who received CT (i.e., matched) fared better than those who received IPT (mismatched). Likewise, among those with negative scores, IPT yielded better outcomes. Their paper was widely cited in the years following its publication, but nearly always in reference to the individual moderator variables rather than to the matching factor and its implications for treatment selection. Perhaps the field was not ready to think in terms of multivariable prediction, but it is also possible that the matching factor idea failed to catch on because individuals' scores on it were not easy to interpret beyond whether they indicated IPT or CT as the preferred treatment. Had the field picked up on the potential for using a version of this approach to advance personalized or precision medicine, we would have been much further along in our understanding of the promise as well as the limitations of this kind of inquiry fifteen years later, when similar efforts were beginning to attract much greater interest and effort from mental health researchers.

One limitation that has come into sharper focus in recent years is the practice of using data from the same participants, without implementing a hold-out or cross-validation approach, for all the steps in the process of generating and testing prescriptive predictions: a)

identifying the variables to be used in the model; b) setting the weights for the model; c) using the model to categorize participants (e.g., as either IPT-preferred and CT-preferred cases); and d) estimating and conducting significance tests on the differences in outcomes between the conditions as a function of the categorization. As has since been well-described in related areas of science, estimates that derive from this process are expected to show high variance and low bias (hallmarks of models that overfit the data) and will therefore be overoptimistic about the utility of the model if it were to be applied in other samples from the same population as the original sample (Hastie, Tibshirani, & Friedman, 2009). More recent efforts in this area, including some of our own, have continued to be subject to some of these same limitations, but we also see progress, as discussed below.

In the first decade of this century, Wolfgang Lutz and colleagues showed how a statistical method developed by avalanche scientists, termed "nearest neighbors," could be applied to the problem of clinical prediction (Lutz et al., 2005). Although most of their work with this method has focused on prognosis – the prediction of outcome in a single context – they also described how it could be used to predict which of two treatments is more likely to benefit a given person (Lutz et al., 2006). The nearest neighbors method works by examining the outcomes of those patients in a sample who are most similar to the target patient (on demographic, symptoms, course, personality, and other measures), for whom the prediction is to be made. The distance between a given patient with known outcome and the target patient is a function of the sum of the squared differences between them, across all baseline variables used for this purpose. To make a prognostic prediction, the average outcome of the target patient's nearest neighbors becomes the estimate of his or her outcome, assuming the same treatment is to be given to that patient. In a prescriptive context, the prediction of the target patient's outcomes derives from a comparison of the average outcome of the nearest neighbors who received treatment X with the average outcomes of nearest neighbors who received treatment Y. Similar to the matching method of Barber and Muenz (1996), the sign of the difference in outcomes between treatments X and Y indicates the preferred treatment for the patient. This method has intuitive appeal in its suggestion that for each patient there is a circumscribed subgroup of other patients who, because they most closely resemble the target patient across the baseline variables used to identify the nearest neighbors, are thus the most relevant for a prediction of their outcome. But the designation of nearest neighbors hinges on many factors and thus requires many decisions before it can be applied, including the selection of the variables to be used to determine near-ness; whether and how differential weights are to be assigned to the variables; and the decision rule to determine how many neighbors' data are used to predict the target patient's outcome.

Developments this decade

In 2013, Helena Kraemer described a novel method for combining moderators of treatment outcome into a combined moderator, which she termed "M*". She argued that given that there rarely are strong moderators in mental health treatment comparison contexts, combining many relatively weak moderator effects could lead to useful differentiations among patients as to which of two treatments is preferable for them. That same year Kraemer and her colleagues demonstrated the method using intake and outcome data obtained in a randomized comparison of an ADM versus Interpersonal Therapy (IPT) for depression (Wallace, Frank, & Kraemer, 2013). They reduced an initial set of 32 baseline variables to 8, using a principal components analysis. Although only one of the variables (psychomotor activation) reached significance as a moderator on its own, all 8 variables were used to construct the M*. When Wallace et al. used the M* to identify patients for whom the suggested treatment was ADM, those randomized to ADM fared better than those who received IPT. The symmetrical effect was observed in those for whom IPT was suggested by the M* implementation. The procedures required to create an M*, which are rather complicated to describe and not closely related to any of the other methods referenced in this paper, can be found in Kraemer (2013) and Wallace et al. (2013).

While Kraemer and her colleagues were developing the M*, our group was experimenting with an approach to treatment selection that yields what we termed the Personalized Advantage Index (PAI; DeRubeis, Cohen et al., 2014). The goal was to work out a method that could be used to estimate, for any patient, the outcomes they would experience in each of two treatments, and to use the difference between those estimates to indicate which treatment is to be preferred, as well as the strength of the indication. In our first demonstration of the approach we used baseline and post-treatment data from our randomized comparison of ADM and CBT to build, for each patient, a model that would be used to make the two predictions for that patient. We elected to demonstrate the approach using linear models with interacting covariates because linear models are common in mental health research. However, machine learning methods can be used to the same purpose, and they have many advantages relative to linear models, such as increasing model prediction accuracy by reducing overfitting (Kuhn & Johnson, 2013). We provided estimates of the advantage of the use of the PAI model, which we believed could be used to anticipate the size of the advantage that could be expected if the algorithm were to determine treatment selection in the same population from which our sample was drawn. On the post-treatment Hamilton Rating Scale for Depression (Hamilton, 1960), the difference in average score between those who received their PAI-indicated treatment and those who did not was approximately 1.8 points, and twice that in the subgroup for whom a relatively large advantage was indicated by the PAI.

The implementations by Wallace et al. (2013) of the M* method and Barber and Muenz's matching method included all participants' data at each step in the procedure, a practice termed "double-dipping." Double-dipping is known to produce models that are overfit to the sample used to develop the model. Overfit models produce over-optimistic predictions of the utility of the application of the model in a different sample from the same population (same clinic, same therapy procedures, etc.), not to mention a different population (e.g., a different clinic, different region of a country). Since the ultimate utility of any modeling procedure derives from its ability to make valid predictions in samples and populations other than the one in which is it developed, it is important to follow the principles that limit or eliminate overfitting. Virtually all modeling efforts in mental health research have, until very recently, used procedures that produce overfit models. In our initial demonstration of the PAI approach, we implemented leave-one-out (LOO) cross validation (Friedman, Hastie, & Tibshirani, 2001), a variant of *k*-fold internal cross-validation, at the weight-setting stage when the coefficients in a linear model are determined. However, the variables in the models were the same for all, having been selected using baseline and outcome information from all the patients. Thus, because our procedure, too, "double-dipped," we have tempered our initial optimism about the applicability of these findings and we recognize that greater care needs to be exercised in the development and within-sample testing of the models, lest we find that they do not travel well, or at all, outside the samples in which they are developed.

Luedtke, Sadikova, & Kessler (2018) have argued that only with sample sizes larger than what is found in almost any randomized trial can we expect to estimate multivariable models that will generalize beyond the dataset in which they are built. In a discussion of their findings from a simulation study, they posit that sample sizes of less than 500 per condition will not support the development of replicable multivariable models. They maintain that precision mental health will progress only if large-scale naturalistic data are used to produce the models. However, there are disadvantages to the use of naturalistic data for model building, and Luedtke et al. acknowledge them. Specifically, when patients are not randomized to treatments, but instead are assigned based on unknown factors, crucial assumptions of model-

building are violated, leading to confounds of unknown magnitude. Moreover, the richness of baseline assessments typical of an RCT is almost always lacking in large-scale naturalistic data.

We agree that where sample sizes are concerned, more is better, especially when models rely on treatment by moderator interaction effects or their machine-learning analogue. However, we believe it is too early to abandon efforts to discover what can be learned using data from randomized trials in which sample sizes are smaller than 500, for two reasons. Firstly, inferences from simulations such as those conducted by Luedtke et al. (2018) are constrained by the values instantiated in the simulations. Although Luedtke et al. might have represented effects as strong as could exist in data gathered in RCTs, it is possible that stronger reliable effects are present in data from at least some randomized trials. Secondly, their simulations did not address the potential for models to guide treatment selection for the subset of patients for whom the model's recommendation is especially strong. It will often be the case that the predicted differential effects of one treatment, relative to another, will be small or modest for a substantial subset of patients, suggesting that either treatment is expected to be about as good (or bad) as the other for those patients. Imagine a crystal ball that can predict with perfect accuracy and precision any patient's outcomes in either of two treatments that are roughly effective on average for patients with the same diagnosis. For some patients it will no doubt predict very similar outcomes irrespective of treatment. Included in this subset would be those for whom the prediction is similarly pessimistic in both treatments, because neither treatment contains the elements required to address the pathology of these patients. It would also include those for whom the prediction is similarly optimistic, either because the patients would improve even without treatment or because they will respond equally well to the specific elements of either treatment. The crystal ball, then, would provide useful recommendations for only those patients who fall into none of the subgroups just described. If the proportion of patients who will do quite well in one treatment and not nearly as well in the other is very low, then even the crystal ball will be of limited utility. In such a case no prescriptive method will be of any detectable use. But if the “little-if-any-difference” subgroups comprised, say, one-half or less of a population, a crystal ball would be very useful indeed. In cases like this, it becomes possible for a modeling effort to produce useful predictions for those it would identify as especially likely to achieve a better response in one treatment, relative to the other. In our work, although we provide results that consider the outcomes of all the patients in a sample, we also show what is observed in the 60% of patients for whom the differential prediction generated by a model is the strongest.

Findings from further work using more refined versions of the methods we implemented in our first effort suggest it is possible that useful models can be produced with “medium-sized” samples. In 2018, we published a paper on the prediction of differential dropout using data from an RCT of prolonged exposure versus cognitive processing therapy for patients with rape-induced PTSD (Keefe, Wiltsey Stirman et al., 2018). In the study from which the data were drawn, dropout rates were high (27% in each condition; Resick, Nishith, Weaver, Astin, & Feuer, 2002). We asked whether, using machine learning methods, we could identify patients whose baseline characteristics predict they would be more likely to drop out of one of the conditions, relative to the other. We employed *k*-fold (in this case 5-fold) internal cross-validation to set the weights used for the individual patient predictions but, as in our initial PAI project , the variable selection stage was not protected from double-dipping. Results were encouraging, with a 21% difference in dropout between those who had been assigned to their PAI-indicated treatment versus their PAI non-indicated treatment, and the difference was even larger (38%) in the portion of the sample with the greatest absolute values on the PAI. We have since re-analyzed these data, using internal cross-validation at both the variable-selection and weight-setting stages (Keefe, Stirman et al., 2018). Although the differences in dropout were not as strong between PAI-indicated and PAI-non-indicated subgroups (20–25% deflated in magnitude), they were nonetheless substantial and significant. We do not know if this is an unusual case, either because random error aligned with the predictions or because this is a population in which the multivariable interactions with these treatments are much stronger than, for example, those instantiated in the simulations implemented by Luedtke et al. (2018). In any event, determination of the potential clinical utility of these approaches will require the additional step of testing and applying the models to a true holdout. This can be done either in a new context with existing data, such as in a similar RCT that has been carried out by other investigators, or in a prospective test designed to estimate the advantage afforded by the implementation of a model in a clinical setting. It will be important to avoid the potential perils of multivariable model building and interpretation before the models are tested for their ability to make valid predictions outside the context in which the models are built. If appropriate protections are not in place during the construction of a model, it very likely will fail in replications and prospective tests.

Some researchers have already reported on their use of machine learning to predict outcomes in true holdout samples, although to date these efforts have focused on prognostic rather than prescriptive predictions. Chekroud et al. (2016) trained machine learning models using 1949 patients from the STAR-D study of treatments for depression (Rush et al., 2008) and found that the prognostic models for some of the treatments predicted outcomes better than chance in a new sample. Specifically, using a model built with data from patients in the citalopram condition of STAR*D, they were able to predict, at better than chance levels, outcomes in the escitalopram plus placebo (n = 151) and the escitalopram plus buproprion (n = 134) conditions in the CO-MED trial (Rush et al., 2011). Other researchers have similarly applied machine learning methods – albeit in smaller samples – to predict treatment outcomes after first episode psychosis (training sample n = 334; test sample n = 108; Koutsouleris et al., 2016) or in response to quetiapine or lithium in bipolar disorder (training sample n = 386; test sample n = 96; Kim et al., 2019).

In medicine more generally, machine learning methods have also been applied in the prediction of heart attacks (e.g., Weiss, Natarajan, Peissig, McCarty, & Page, 2012), live birth following IVF (Qiu, Li, Dong, Xin, & Tan, 2019), and cancer (for a review see Kourou, Exarchos, Exarchos, Karamouzis, & Fotiadis, 2015). Further, a meta-analysis by Lee et al. (2018) has found that machine learning methods using neuroimaging, phenomenological, and genetic data are able to predict therapeutic outcomes in mood disorders, but again these were predictions within a given treatment, not differential predictions between two or more treatments. These findings indicate that even without very large samples, prognostic models derived using machine learning methods have the potential to be quite powerful.

As suggested by Kessler et al. (2018), given the notoriously low power of interaction terms in multivariable models, a promising approach to precision mental health may include the construction of prognostic models for each of the individual treatments of interest. Comparisons of an individual's prognosis for a given treatment with those for other treatments would then be used to guide treatment decisions. If the prognoses are similar, other considerations, such as cost or patient preference, will weigh more heavily. If, however, the prognoses estimated for the different treatments are quite discrepant, this information could be used to inform treatment decisions.

Successful tests of prospective differential predictions between two or more treatments, developed with linear models or machine-learning based multivariable algorithms, are the necessary next steps in this area of research. We are not aware of any completed or ongoing tests that meet these criteria, but we know of two groups who are attempting to secure funding for this kind of work. We eagerly await the initiation of such prospective tests and their findings, which will provide important clues as to whether actuarial-based decision support tools have a future in precision mental health.

Although it is premature to predict with confidence that actuarial

models of the kinds referred to in this paper will play a substantial role in mental health practice, it is important to think about what that role might be. The clinic of the future could implement the predictions of models to the exclusion of all other considerations, but that is as unlikely as it is unwise. Availability of the treatments in question, their cost, and other practical matters will no doubt always factor into treatment decisions. Another important input is the patient's preference. A treatment selection system will likely work best if it is implemented in the context of joint decision-making between the patient and provider, rather than because the "computer says so." Similarly, providers whose recommendations are ignored and overridden by an actuarial-based system will not be enthusiastic about the system and likely will resist its implementation. Thus, the recommendations that will flow from an actuarial method will survive and contribute to mental health systems if they are treated as a source of information rather than a dictate. A medication-averse patient with MDD who learns that previous patients whose course, pattern of symptoms, etc. have responded better to ADMs than to CBT might become more open to a discussion of medications as an option for them. Similarly, a therapist who believes that a patient should be provided a course of prolonged exposure to address their symptoms of PTSD might take the output of a multivariable algorithm into account if it predicts that the patient is less likely to drop out of cognitive processing therapy than prolonged exposure treatment.

Tests of the potential utility of any treatment selection method will involve contrasts of outcomes experienced by patients who received the treatment selected for them versus outcomes observed in a comparison group. The most common comparison in studies to date, using data from already-completed RCTs, have contrasted those who received the indicated treatment with those who received the non-indicated treatment (Cohen & DeRubeis, 2018). This contrast highlights maximally the potential for a method to distinguish subgroups of patients who will show different patterns of response to the two treatments. However, the scenario it represents has no analogue in clinical practice, since it would require the application of the algorithm and then a decision to act directly against the information it provides. A more stringent and therefore more informative contrast is of algorithm-indicated assignment versus random assignment. If two treatments produce roughly equal average outcomes, a selection algorithm will likely assign about one-half of the patients to each of two treatments. In this case, the magnitude of the difference in outcomes between patients in the algorithm-indicated group and those assigned randomly will approximate one-half of the difference observed in the indicated vs. non-indicated comparison. This is because about half of those randomly assigned will be given the algorithm-indicated treatment, by chance.

Comparisons with other assignment regimes would provide information that is directly relevant to clinical application. If one of the two treatments is, on average, superior to the other, it still may be that a subgroup could be identified, using statistical methods, that would be expected to benefit more from the inferior treatment. In this case, a crucial comparison is between the average expected benefit of algorithm assignment and the expected benefit of assigning every patient to the superior treatment. Another clinically-applicable comparison is between algorithmic assignment and an "assignment as usual" or "clinical judgment" regime. If the algorithm outperforms clinical judgment, a result that would not surprise researchers who have compared clinical versus actuarial prediction (Dawes, Faust, & Meehl, 1989), then its application could improve the overall effectiveness and efficiency of the system in which it is used. If the algorithm simply replicates what clinicians would decide, there is no expected advantage from its application, although a peek inside the algorithm could inform an understanding of how clinicians combine information to make their decisions. If clinical judgment were to outperform an algorithm, it would motivate an attempt to understand what information the clinicians were using and how they were combining it. Perhaps the best performers among the clinicians could be identified, and others could be trained to emulate their performance.

Estimates of the expected benefits of the use of algorithmic-indicated assignment relative to non-indicated assignment, random assignment, or to assignment of all cases to the superior treatment can be produced using data from any randomized trial, including trials that have already been completed. Moreover, if a randomized clinical trial includes an effort to determine what assignment would have been made for each patient if it were not random, the algorithm can be compared to clinical judgement or "assignment as usual" as well. An example of this comes from a randomized trial of CBT versus IPT for outpatients presenting for depression (Lemmens et al., 2015). Huibers et al. (2015) had implemented the PAI approach and found that patients who were assigned to the algorithm-indicated treatment fared better than those assigned to their non-indicated treatment. When the investigators compared the recommendations produced by the PAI with those made by a panel of judges, they found that the PAI-based recommendations were associated with better outcomes, on average (Van Bronswijk, Lemmens, Huibers, & Peeters, 2019). Koutsouleris et al. (2018) also have demonstrated that a machine learning algorithm can outperform clinical judgment, in this case in the prediction of clinical risk in a mixed sample of patients with recent-onset MDD and recent-onset psychosis. These investigators are to be commended for having had the foresight to obtain clinicians' judgments, which allow for this important kind of comparison.

Summary and conclusions

Investigations of person characteristics that can predict treatment response, including differential response to alternative evidence-based interventions, have become increasingly sophisticated over the past 50 years. In the past decade clinical scientists have begun to design randomized intervention studies and implement statistical analyses that anticipate the use of the resulting algorithms in applied settings. It is too soon to tell whether and to what extent these efforts will address the question "What works for whom?" with sufficient precision and power to improve predictions of treatment response. It may be that multivariable machine-learning-based algorithms require sample sizes much larger than has been the norm in randomized trials if they are to generalize beyond the specific sample. Thus it will be important for research that uses information from large databases to progress at the same time that the limits of the use of randomized data are being tested.

Whether from randomized trials or big data, the further evolution of research in this area will require prospective tests of robust treatment selection systems. If these tests suggest that appreciable increments in the efficiency and effectiveness of mental health interventions can be realized, it will still require intelligent implementation that takes account of the preferences and goals of clinician and client stakeholders. This is an exciting time for research that takes up the challenge and attempts to realize the promise of actuarial methods, prophesied long ago by Paul Meehl (1954), one of the great thinkers in the history of clinical psychology.

Acknowledgements

The author wishes to thank Colin Xu for his help with reviewing and editing earlier versions of this manuscript.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.brat.2019.103506.

References

Barber, J. P., & Muenz, L. R. (1996). The role of avoidance and obsessiveness in matching patients to cognitive and interpersonal psychotherapy: Empirical findings from the

treatment for depression collaborative research program. *Journal of Consulting and Clinical Psychology, 64*(5), 951–958. Retrieved from https://www.ncbi.nlm.nih.gov/pubmed/8916624.

Chekroud, A. M., Zotti, R. J., Shehzad, Z., Gueorguieva, R., Johnson, M. K., Trivedi, M. H., et al. (2016). Cross-trial prediction of treatment outcome in depression: A machine learning approach. *Lancet Psychiatry, 3*(3), 243–250. https://doi.org/10.1016/S2215-0366(15)00471-X.

Cohen, Z. D., & DeRubeis, R. J. (2018). Treatment selection in depression. *Annual Review of Clinical Psychology, 14*, 209–236. https://doi.org/10.1146/annurev-clinpsy-050817-084746.

Dawes, R. M., Faust, D., & Meehl, P. E. (1989). Clinical versus actuarial judgment. *Science, 243*(4899), 1668–1674. Retrieved from https://www.ncbi.nlm.nih.gov/pubmed/2648573.

DeRubeis, R. J., Cohen, Z. D., Forand, N. R., Fournier, J. C., Gelfand, L. A., & Lorenzo-Luaces, L. (2014a). The personalized advantage Index: Translating research on prediction into individualized treatment recommendations. A demonstration. *PLoS One, 9*(1), e83875. https://doi.org/10.1371/journal.pone.0083875.

DeRubeis, R. J., Gelfand, L. A., German, R. E., Fournier, J. C., & Forand, N. R. (2014b). Understanding processes of change: How some patients reveal more than others-and some groups of therapists less-about what matters in psychotherapy. *Psychotherapy Research, 24*(3), 419–428. https://doi.org/10.1080/10503307.2013.838654.

DeRubeis, R. J., Hollon, S. D., Amsterdam, J. D., Shelton, R. C., Young, P. R., Salomon, R. M., et al. (2005). Cognitive therapy vs medications in the treatment of moderate to severe depression. *Archives of General Psychiatry, 62*(4), 409–416. https://doi.org/10.1001/archpsyc.62.4.409.

Elkin, I., Parloff, M. B., Hadley, S. W., & Autry, J. H. (1985). NIMH treatment of depression collaborative research program. Background and research plan. *Archives of General Psychiatry, 42*(3), 305–316. Retrieved from https://www.ncbi.nlm.nih.gov/pubmed/2983631.

Fournier, J. C., DeRubeis, R. J., Shelton, R. C., Gallop, R., Amsterdam, J. D., & Hollon, S. D. (2008). Antidepressant medications v. cognitive therapy in people with depression with or without personality disorder. *British Journal of Psychiatry, 192*(2), 124–129. https://doi.org/10.1192/bjp.bp.107.037234.

Fournier, J. C., DeRubeis, R. J., Shelton, R. C., Hollon, S. D., Amsterdam, J. D., & Gallop, R. (2009). Prediction of response to medication and cognitive therapy in the treatment of moderate to severe depression. *Journal of Consulting and Clinical Psychology, 77*(4), 775–787. https://doi.org/10.1037/a0015401.

Friedman, J., Hastie, T., & Tibshirani, R. (2001). *The elements of statistical learning, Vol. 1*. New York: Springer series in statistics.

Hamilton, M. (1960). A rating scale for depression. *Journal of Neurology, Neurosurgery & Psychiatry, 23*, 56–62. https://doi.org/10.1136/jnnp.23.1.56.

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The elements of statistical learning: Prediction, inference and data mining*. New York: Springer-Verlag.

Huibers, M. J., Cohen, Z. D., Lemmens, L. H., Arntz, A., Peeters, F. P., Cuijpers, P., et al. (2015). Predicting optimal outcomes in cognitive therapy or interpersonal psychotherapy for depressed individuals using the personalized advantage Index approach. *PLoS One, 10*(11), e0140771. https://doi.org/10.1371/journal.pone.0140771.

Keefe, J. R., Stirman, S., Cohen, Z., DeRubeis, R., Smith, B., & Resick, P. (2018a). What works for whom" in sexual trauma PTSD. Paper presented at the association for psychological science, San Francisco, CA.

Keefe, J. R., Wiltsey Stirman, S., Cohen, Z. D., DeRubeis, R. J., Smith, B. N., & Resick, P. A. (2018b). In rape trauma PTSD, patient characteristics indicate which trauma-focused treatment they are most likely to complete. *Depression and Anxiety, 35*(4), 330–338. https://doi.org/10.1002/da.22731.

Kessler, R. C. (2018). The potential of predictive analytics to provide clinical decision support in depression treatment planning. *Current Opinion in Psychiatry, 31*(1), 32–39. https://doi.org/10.1097/YCO.0000000000000377.

Kessler, R. C., van Loo, H. M., Wardenaar, K. J., Bossarte, R. M., Brenner, L. A., Ebert, D. D., et al. (2017). Using patient self-reports to study heterogeneity of treatment effects in major depressive disorder. *Epidemiology and Psychiatric Sciences, 26*(1), 22–36. https://doi.org/10.1017/S2045796016000020.

Kim, T. T., Dufour, S., Xu, C., Cohen, Z. D., Sylvia, L., Deckersbach, T., et al. (2019). Predictive modeling for response to lithium and quetiapine in bipolar disorder. *Bipolar Disorders*. https://doi.org/10.1111/bdi.12752.

Kourou, K., Exarchos, T. P., Exarchos, K. P., Karamouzis, M. V., & Fotiadis, D. I. (2015). Machine learning applications in cancer prognosis and prediction. *Computational and Structural Biotechnology Journal, 13*, 8–17. https://doi.org/10.1016/j.csbj.2014.11.005.

Koutsouleris, N., Kahn, R. S., Chekroud, A. M., Leucht, S., Falkai, P., Wobrock, T., et al. (2016). Multisite prediction of 4-week and 52-week treatment outcomes in patients with first-episode psychosis: A machine learning approach. *Lancet Psychiatry, 3*(10), 935–946. https://doi.org/10.1016/S2215-0366(16)30171-7.

Koutsouleris, N., Kambeitz-Ilankovic, L., Ruhrmann, S., Rosen, M., Ruef, A., Dwyer, D. B., et al. (2018). 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.

Kraemer, H. C. (2013). Discovering, comparing, and combining moderators of treatment on outcome after randomized clinical trials: A parametric approach. *Statistics in Medicine, 32*(11), 1964–1973. https://doi.org/10.1002/sim.5734.

Kuhn, M., & Johnson, K. (2013). *Applied predictive modeling26*. New York: Springer.

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.

Lemmens, L. H., Arntz, A., Peeters, F., Hollon, S. D., Roefs, A., & Huibers, M. J. (2015). Clinical effectiveness of cognitive therapy v. interpersonal psychotherapy for depression: Results of a randomized controlled trial. *Psychological Medicine, 45*(10), 2095–2110. https://doi.org/10.1017/S0033291715000033.

Leykin, Y., DeRubeis, R. J., Gallop, R., Amsterdam, J. D., Shelton, R. C., & Hollon, S. D. (2007). The relation of patients' treatment preferences to outcome in a randomized clinical trial. *Behavior Therapy, 38*(3), 209–217. https://doi.org/10.1016/j.beth.2006.08.002.

Luedtke, A., Sadikova, E., & Kessler, R. C. (2018). Sample size requirements for multivariate models to predict between-patient differences in best treatments of major depressive disorder. *Clinical Psychological Science, 7*(3), 445–461.

Lutz, W., Leach, C., Barkham, M., Lucock, M., Stiles, W. B., Evans, C., et al. (2005). Predicting change for individual psychotherapy clients on the basis of their nearest neighbors. *Journal of Consulting and Clinical Psychology, 73*(5), 904–913. https://doi.org/10.1037/0022-006X.73.5.904.

Lutz, W., Saunders, S. M., Leon, S. C., Martinovich, Z., Kosfelder, J., Schulte, D., et al. (2006). Empirically and clinically useful decision making in psychotherapy: Differential predictions with treatment response models. *Psychological Assessment, 18*(2), 133–141. https://doi.org/10.1037/1040-3590.18.2.133.

Meehl, P. E. (1954). *Clinical versus statistical prediction: A theoretical analysis and a review of the evidence*. Minneapolis, MN, US: University of Minnesota Press.

Michel, J. B., Shen, Y. K., Aiden, A. P., Veres, A., Gray, M. K., Google Books, T., et al. (2011). Quantitative analysis of culture using millions of digitized books. *Science, 331*(6014), 176–182. https://doi.org/10.1126/science.1199644.

Paul, G. L. (1967). Strategy of outcome research in psychotherapy. *Journal of Consulting Psychology, 31*(2), 109–118. Retrieved from https://www.ncbi.nlm.nih.gov/pubmed/5342732.

Project MATCH Research Group (1998). Matching alcoholism treatments to client heterogeneity: Treatment main effects and matching effects on drinking during treatment. *Journal of Studies on Alcohol, 59*(6), 631–639. https://doi.org/10.15288/jsa.1998.59.631.

Qiu, J., Li, P., Dong, M., Xin, X., & Tan, J. (2019). Personalized prediction of live birth prior to the first in vitro fertilization treatment: A machine learning method. *Journal of Translational Medicine, 17*(1), 317. https://doi.org/10.1186/s12967-019-2062-5.

Resick, P. A., Nishith, P., Weaver, T. L., Astin, M. C., & Feuer, C. A. (2002). A comparison of cognitive-processing therapy with prolonged exposure and a waiting condition for the treatment of chronic posttraumatic stress disorder in female rape victims. *Journal of Consulting and Clinical Psychology, 70*(4), 867–879. Retrieved from https://www.ncbi.nlm.nih.gov/pubmed/12182270.

Rush, A. J., Trivedi, M. H., Stewart, J. W., Nierenberg, A. A., Fava, M., Kurian, B. T., et al. (2011). Combining medications to enhance depression outcomes (CO-MED): Acute and long-term outcomes of a single-blind randomized study. *American Journal of Psychiatry, 168*(7), 689–701. https://doi.org/10.1176/appi.ajp.2011.10111645.

Rush, A. J., Wisniewski, S. R., Warden, D., Luther, J. F., Davis, L. L., Fava, M., et al. (2008). Selecting among second-step antidepressant medication monotherapies: Predictive value of clinical, demographic, or first-step treatment features. *Archives of General Psychiatry, 65*(8), 870–880. https://doi.org/10.1001/archpsyc.65.8.870.

Van Bronswijk, S. C., Lemmens, L. H. J. M., Huibers, M. J. H., & Peeters, F. P. M. L. (2019). *Treatment selection for depression: Therapist recommendations or machine learning predictions?* (Unpublished manuscript).

Wallace, M. L., Frank, E., & Kraemer, H. C. (2013). A novel approach for developing and interpreting treatment moderator profiles in randomized clinical trials. *JAMA Psychiatry, 70*(11), 1241–1247. https://doi.org/10.1001/jamapsychiatry.2013.1960.

Weiss, J. C., Natarajan, S., Peissig, P. L., McCarty, C. A., & Page, D. (2012). Machine learning for personalized medicine: Predicting primary myocardial infarction from electronic health records. *AI Magazine, 33*(4) 33-33.