MATCH: Client-therapist matching with machine learning

Seth M. Peacock, Ian H. Goodwin, Ryan K. Wood, Connor J. McBride, Ammon C. Brock, David M. Erekson, Zachary M. Boyd

Brigham Young University, Provo, UT, USA

Abstract

A good match between clinicians and clients can substantially impact psychotherapy outcomes. No proposed matching methodology, however, accounts for the multitude of relevant variables, such as complaint type, demographics, life experiences, and personality, as well as for practical concerns such as therapist availability. Prior work has often focused on single-variable models and average effects across populations, and studies that take more multifaceted approaches base their results on simulations or carefully constructed situations. To nuance our matching, we estimated the complex effect of therapists on client outcomes using machine learning trained on high dimensional data from CCAPS and OQ-45 surveys from 2014 to 2019. We used these predictions to produce constrained matches, optimizing outcomes for cohorts of clients. Using our method, called Matching Assistant for Therapists and Client Health (MATCH), clients assigned would experience better outcomes on average with minimal impact on their wait-time and few administrative changes.

Numerous studies of psychotherapy have found that data-driven matching of clients to therapists can improve client outcomes, either using machine learning (ML)[1,2] or other data-informed methods.[3–5] Good client-clinician matching promises to lead to better therapeutic alliance, improved client functioning, reliable decrease in depressive symptoms, and positive change in many other client outcomes. Implementing data-driven client-clinician matching in practice could be highly beneficial, but attempts to do so immediately face the problem that client-

clinician matching is highly *constrained* by practicalities, such as the number of clients that a favored therapist can see at once or the related issue of client wait time. Past simulation studies and clinical trials have generally not accounted for this, for example by working in controlled settings where each client could meet with the therapists that researchers designated. The present study (1) quantifies the significant bottleneck that can arise in practice in trying to use ML recommendations to match clients to clinicians, (2) proposes a solution to this problem by casting the constraints in a classical computer science framework, and (3) reports how much of the matching intervention effect can still be achieved in practice using our framework.

However constrained, client-clinician matching aims to improve client outcomes by leveraging what are called *therapist effects*. The therapist effect is the difference in client outcomes (such as drop-out, questionnaire scores, etc.) due to the specific therapist that the client was paired with, compared to the average outcome for that client across all therapists.[6] The therapist effect is related to a wide range of factors, including therapist experience level[7] and treatment modality.[8] A therapist may also excel at certain types of concerns,[9] and clients may do well with therapists that are similar to themselves.[8] Moreover, choice of therapist can have a greater effect on client outcomes than choice of treatment.[10] Still, the exact reasons for the difference in client outcomes between therapists are unclear,[11] and there is a need for client-centric approaches that integrate many variables to compute optimal matches, as opposed to variable-centric studies that emphasize the average population-level effect of individual factors.

A common method of matching is to assign a client to the next available therapist in a given clinic—effectively random assignment. This is the current and historical approach of Brigham Young University's (BYU's) Counseling and Psychological Services (the clinic at the authors' home institution), but elsewhere, researchers have tried to account for therapist effects by using other matching methods, with promising results. For example, Boswell et al. found that matching clients with therapists who had higher historical performance in the client's area of concern yielded better clinical results.[3] We note, however, that the randomized trial they conducted testing these matches was not employed clinic-wide but on select clients.

While top-concern matching has been empirically verified, ML models are no less suited for the task of matching. They are a versatile tool in psychotherapy with a rapidly increasing range of applications. Researchers have used ML to diagnose clients,[12] anticipate client drop-out,[13] predict

treatment outcomes,[14] flag suicidal ideation,[12] give therapy-approach recommendations,[15] and encode client behavior.[16] ML has been successfully applied to a very broad range of problems that can be formulated in terms of mathematical optimization.[17] The models take in many variables and find complex interactions and non-linear relationships that a simple one-variable analysis, like top concern, might not uncover. Moreover, ML models can produce quantified predictions, simplifying the comparison between counterfactual situations and informing action. These quantified predictions make possible the method that we put forward in this paper for addressing client-therapist pairings in the presence of constraints. Using ML models paired with appropriately randomized data, we can have better-than-chance estimates even when the underlying causal effects are not fully understood.

Our work builds on several prior studies that have successfully used ML in client-therapist matching. Boswell et al. employed ML models to analyze past client records and estimate each therapist's effect on an outcome called Treatment Outcome Package (TOP) in different areas for each therapist. They then validated in a clinical trial that matching a client to a therapist that had a record of doing well with the client's top concern improved client outcomes in TOP relative to random matching. Delgadillo et al. trained a decision tree to predict client outcomes using ML, then used the tree to group therapists who performed similarly. They found in simulations that matching clients to those groups resulted in significant outcome improvement (as measured by score on the PHQ-9, or Patient Health Questionnaire-9).[1] Bronswijk et al. used ML to predict which treatment modality (CBT vs IPT) would best reduce a patient's Beck Depression Inventory–version II score and validated the ML-based matches in two clinical trials.[18] While these studies used ML in a similar way as our research, they do not address constrained matching as we do. We also have nearly double the sample size of the previous largest study, included all therapists at the clinic who did individual therapy in a normally operating clinical setting, and focused on matching to individual therapists rather than groups or types of therapists. These differences allow us to make predictions specific to the nuances of a given set of therapists and clients and make it more reasonable and practical to apply those predictions in clinical practice.

Moving from purely predictive models and controlled clinical trials to application requires additional considerations. For example, a client might match with a therapist who already has a full caseload. Should the client wait or visit some other therapist? This concern is especially

important in the common situation where the ML model matches a disproportionately large number of clients with a therapist whose average effectiveness (however it is measured) is higher than average (as happens in Fig. 3 herein). In other fields, researchers have addressed analogous constrained matching situations, such as organ donation,[19] consumers partnering with manufacturers,[20] and assigning students to dorms.[21] One way to match is to use queues, partitioning one of the sets to be matched in lists according to some criteria, with a possibility of switching lists; another is to consider a fixed number of items or parties to be matched at a time, and then create pairs to optimize an outcome or to ensure that no two parties would prefer to switch.[22] We chose to group clients together and then optimize for the total improvement of the group, as this ensures that each group of clients has optimal expected outcomes as a whole (see Fig. 1). While much research has gone into finding effective constrained matching algorithms in other fields, we know of none that look at constrained matching algorithms for psychotherapy.

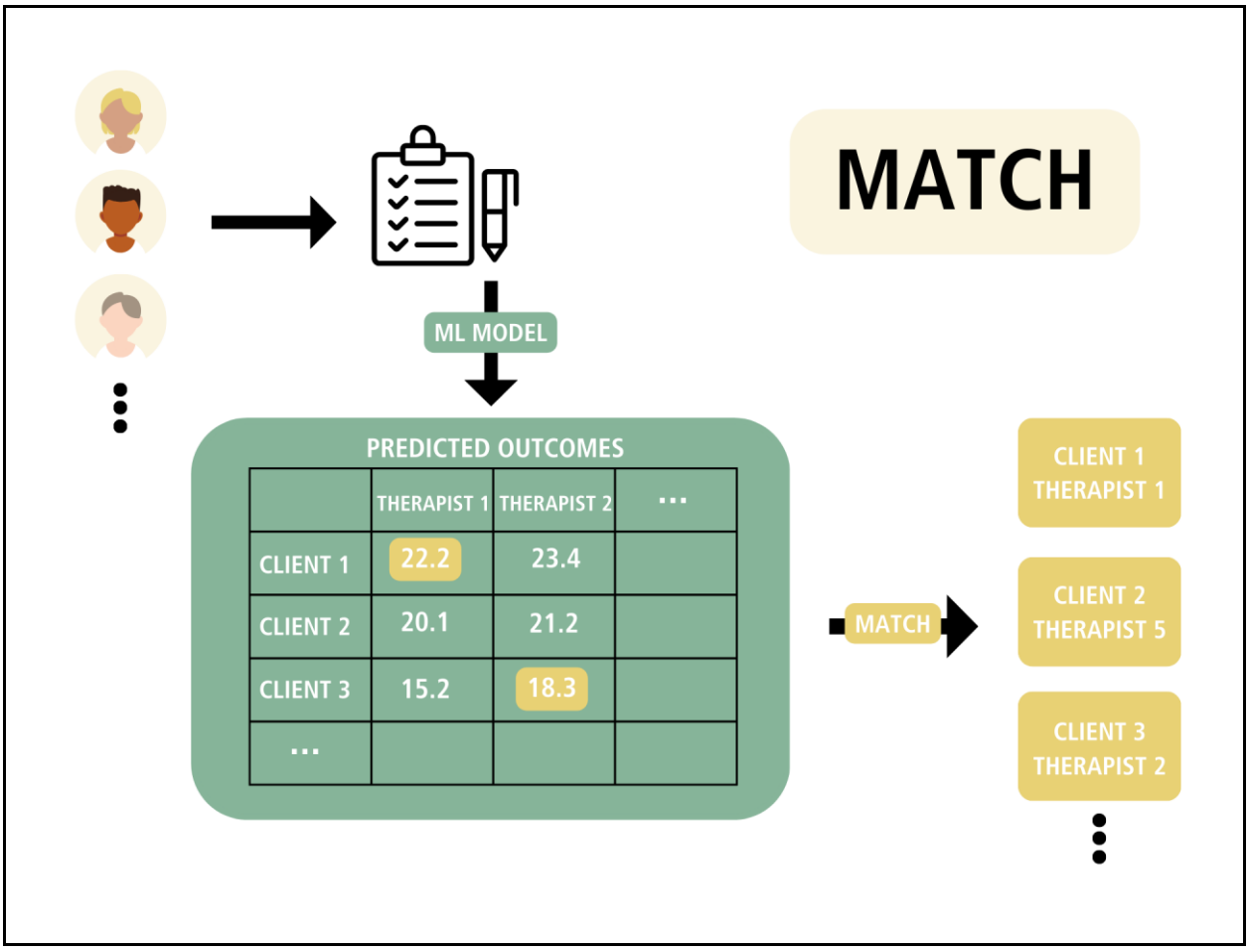

Figure 1: MATCH. This is our algorithm. We take use intake surveys from *n* clients and match them with therapists, who together have *n* openings. First, using an ML model, we predict the outcome of each client with each therapist, then select the pairings that result in the highest total (predicted) improvement in outcome for the cohort of clients.

We draw upon a dataset of 8,541 clients who were primarily college students at BYU, including demographic information, Counseling Center Assessment of Psychological Symptoms-62 (CCAPS-62), and Outcome Questionnaire 45.2 (OQ-45). CCAPS-62 was designed for college students, to assess their presenting-concerns.[14] OQ-45 was designed to assess the effectiveness of psychotherapy throughout a course of therapy.[15] After assessing a variety of models, we discuss which features proved to have the most predictive power and report the prediction error of the best-fit model, an extra trees regressor. Finally, we offer a solution to two practical obstacles to implementing our model in a clinical setting: therapist availability and continuous client intake. This solution, together with our ML model, comprises an algorithm we call the Matching Assistant for Therapists and Client Health (MATCH–see Fig. 1). We show that incorporating realistic scheduling constraints by combining a predictive model with algorithmic constrained matching can provide one-third of the benefit of (less realistic) unconstrained matching. Matching is not only theoretically beneficial but can improve client outcomes without much noticeable change in the normal client intake process.

Table 1: Results. The average improvement in outcomes for clients between 2014 and 2018 is 7.4 OQ-45 points. If each client could see their top-predicted therapist, we predict we would see a further increase of 3 OQ points. Furthermore, we estimated using a retrospective comparison (see Methods) that the effect of a client meeting with one of their top ten therapists was between 0.017 to 3.67 OQ points of improvement on average. With constraints on therapist availability, our algorithm, MATCH, is estimated to yield nearly one third of the benefit of unconstrained matching, considering cohorts of 100 clients (see Results and Methods). The model explained 8.8% of the variability in client outcomes and had a mean absolute error (MAE) of 12.64 OQ-45 points.

	Random assignment	**Matching with unlimited capacity**	**MATCH (cohort size 100)**
Average Benefit (OQ-45 drop)	7.4	10.4	8.3

ML Model R Squared	ML Model MAE	MATCH Effect Size (Top 10)
0.088	12.64 (13.23 baseline)	1.84 (CI: 0.017-3.67)

Results

ML predictions better than random chance

We trained numerous machine learning models to predict the net difference in OQ-45 score across a course of therapy, first score minus last, using a method that vigorously corrects for multiple comparisons (see Methods and Supplement). The best-predicting model, as measured by mean absolute error (MAE) on the test data, was an extremely randomized trees model or extra trees. Extra trees is an altered version of random forests that randomly selects, in each decision tree, both which feature to consider and where to split on the given feature. It had, on a held-out set (never used to choose parameters or hyperparameters), a mean prediction error of 12.67 (compared to a baseline 13.23 when all clients are predicted to have the average outcome) and R-squared of 8.8%. A one-sided permutation test (see Methods) showed that the error of our model was better than random chance ($p < 0.01$).

A variety of features found to be predictive

Before selecting a model, we selected features. Some features we ignored in the cleaning process (see Methods) and on the rest we used Lasso feature selection to produce our final feature set. Table 2 gives the selected features, ordered according to feature importance as determined by our final tree-based model, and some features that were excluded by Lasso. Importance decreases from top to bottom and from left to right. These features do not include any demographic information, suggesting that these features do not predict a client's drop in OQ-45 as well as other features.

Predictive Features		
Intake OQ-45 Score	"I have sleep difficulties."*	"I don't enjoy being around

		people as much as I used to.”*
“I become anxious when I have to speak in front of audiences.”*	Have you had prior counseling?	“I feel confident that I can succeed academically.”*
“I feel helpless.”*	How much distress do physical health problems cause you?	“I am unable to keep up with my schoolwork.”*
“I am easily frightened or startled.”*	“I feel irritable.”*	“I make friends easily.”*
“I feel uncomfortable around people I don’t know.”*	“I feel disconnected from myself.”*	“Confusion about religious beliefs or values”
“I lose touch with reality.”*	“I am concerned that other people do not like me.”*	“My family is basically a happy one.”*
“My heart races for no good reason.”*	"I get the emotional help and support I need from my family."	“I am satisfied with my body shape.”*
“I feel self conscious around others.”*	"I get the emotional help and support I need from my social network (e.g., friends & acquaintances)."	“The less I eat, the better I feel about myself.”*
“I feel comfortable around other people.”*	“I feel tense.”*	How often do you feel stressed about your financial situation right now?
“I enjoy my classes.”*	“I have difficulty controlling my temper.”*	“I am anxious that I might have a panic attack while in public.”*
“I think about food more than I would like to.”*	Harassment/Abuse (How many)	“I feel that my family loves me.”*
Self-injury (how many)	How much distress do sexual concerns cause you?	Traumatic experiences (how many)
“I diet frequently.”*	Therapist ID	How much distress does

		pornography cause you?
To what extent does religion or spirituality play an important role in your life?	Disabilities (how many)	How much distress does your sexual orientation or identity cause you?
Example Exclusions by Lasso		
Demographics	**Concerns**	**History**
Gender	Drugs and Alcohol	Prior medication
Race	Perfectionism	Prior hospitalization
Religion	Anger	
Age	Marriage/relationship	
Homelessness		

Table 2: Final features. This is a list of features that Lasso selected for use in the extra trees model and some features excluded by Lasso. The importance, given by our final model, extra trees, of each feature decreases from top to bottom and from left to right. The items marked with an asterisk (*) come from CCAPS-62. All the features are on a 5-point scale–either Likert or similar–besides intake OQ-45, Therapist ID, prior counseling, and the "how many" features. We note that our final feature set included none of the clients' demographics. This suggests that the client's concern better predicts the change in OQ-45 better than their demographic. Besides excluding the demographics, Lasso left out questions about the client's past psychiatric treatment as well as some questions about the client's concerns and past history. While an extreme value on the OQ-45 is likely to be followed by a less extreme value (regression to the mean), this does not present a concern for our use since we are comparing differences in OQ relative to other choices in therapists.

ML pairings yield better than random outcomes for clients

Using the model predictions, we estimated, via simulation, the effect of each client meeting with their optimal therapist (more precisely, the one that yielded the best predicted drop in OQ-45). When optimally matched, the clients had a predicted mean that was 3.57 OQ-45 points higher

than the historical average. We also estimated, using observed outcomes, the effect of near-optimal matching by looking at clients in a held-out dataset (never trained or tested on) who met with one of their top-ten therapists. We are 95% confident that a client meeting with one of a personalized top-ten would improve between 0.017 to 3.67 OQ-45 points more than that client would with random matching (see Fig. 2b). We also estimated the proportions of clients who would improve by at least 14 OQ-45 points, worsen by at least 14 points, and improve at all. The number 14 was chosen because a difference of this magnitude is the threshold for clinically significant change on the OQ-45.[23] With random matching, 32% of courses of therapy resulted in a reduction of 14 points or more. Using noise-injected estimates to reduce bias, we estimate that this proportion increases to 43% when clients are paired with their optimal therapist. We also expect an additional 4% of clients to show some reduction in OQ-45, and 1% fewer who will significantly deteriorate (meaning their OQ-45 score increases at least 14 points). (See Fig. 2a.)

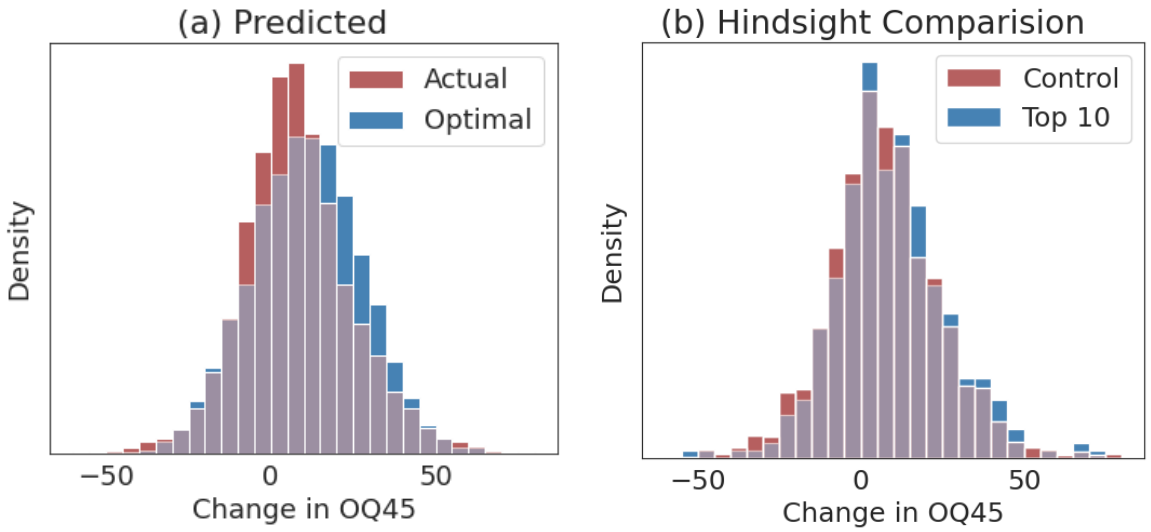

Figure 2: Comparing scores for intervention and control.

Figure 2a: Predicted optimal and actual. For the optimally matched distribution, 74% had a lower OQ-45 after therapy than before, 43% decreased by a clinically significant amount, and 7% increased by a clinically significant amount. Those same numbers for the actual outcomes are 69%, 32%, and 8%. This means that the intervention could increase the number of those who have an improvement as measured in OQ-45 and decrease the proportion that would see a deterioration.

Figure 2b: Hindsight comparison of top-ten to random-assignment. In a held-out historical data set, we predicted the top ten therapists for each client, and then marked the client as treated if they actually met with one of those therapists. The scores for the treatment group improved more on average than an independent control group, improving by 1.86 OQ-45 points more after therapy than the random matches (95% CI, 0.02 to 3.67).

Therapist capacity constraints strongly impact matching feasibility

Across all clients in our training/test data, we found the proportion of times that each therapist was chosen as the optimal and was in a client's top-ten (see Fig. 3).

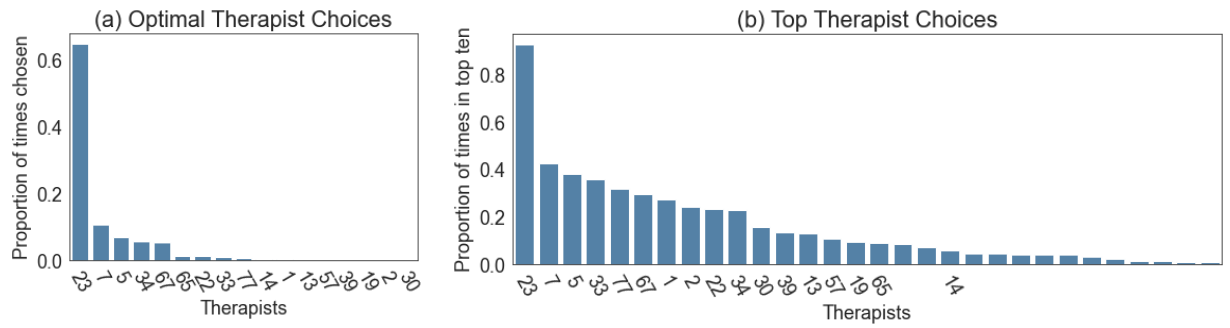

Figure 3: Therapist assignment with unconstrained matching.

Figure 3a: Optimal therapist. This graph shows the proportion of times that a therapist was the optimal therapist for each client in our data. One therapist was the optimal far more often than the others.

Figure 3b: Top ten therapists. This graph shows how often each therapist appears in a clients' top ten therapist choices, across all clients. The same therapists that were in 2a appear in this figure and are numbered in both. Note that the ordering is different and that the proportions are more leveled out. Nonetheless, we need to address the disproportionate assignment. We did so by optimizing matches for a collection of clients and therapists simultaneously. Since we can adjust according to therapist availability, no therapist would become overbooked.

We saw that one particular therapist was the projected optimal therapist in over 60% of cases. This presents a problem for implementing predicted best matches, since they would overload one therapist. Looking at the top ten, we see that the skewness of the distribution diminishes (see Fig. 3b), but not enough to eliminate the need for intentional load-balancing between therapists while implementing a matching protocol.

Appropriately constrained matching with MATCH still benefits clients

We then considered how to optimize client outcomes while respecting therapists' availability. Not every patient can be matched with their optimal therapist or even one of their top ten best matches, but we can still improve client outcomes given the constraints. We used the Hungarian Algorithm to maximize the total change in OQ-45 score for cohorts of clients. Specifically, we take the first *n* clients to come on the waiting list and match them to the next *n* therapist openings in such a way that gives the maximum mean OQ-45 improvement. The Hungarian Algorithm

finds the pairings that give the maximum predicted mean improvement, given the predicted mean improvement for every possible pairing between the two groups. We estimated the average difference in OQ-45 for clients in these groups, varying the group size from 1 to 1,000 and estimating the number of available appointments using historical data. We found that this method yields almost one third of the effect seen in the absence of capacity constraints on matching (see Fig. 4). With batch sizes of 50 and 200, we would expect that clients would have a drop of 8.27 and 8.35 OQ-45 points, respectively, after therapy. At a group size of 90 the average outcome has reached 90% of the improvement achievable at much larger group sizes. We note these sizes specifically because a group size of 50 corresponds to a normal week at CAPS, and 200 corresponds to the first two weeks of a semester.

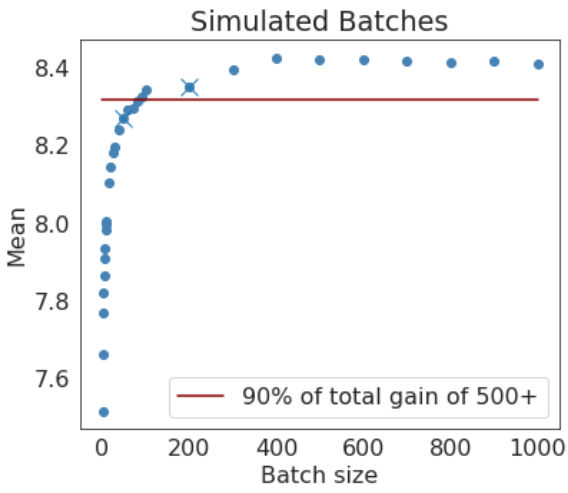

Figure 4: Average OQ-45 improvement as a function of group sizes. We simulated matching cohorts of clients. Note that about 90% of the benefit of the large (500+) group sizes occurs at a group of about 100. CAPS starts about 50 new courses of therapy during a normal week and 200 at the start of the semester; our estimated average outcomes for those group sizes are 8.27 and 8.35. We denoted these points with overlaid x's on the graph. This means that we can have most of the benefit of our matching using a realistic number of known therapist openings.

Discussion

Our results demonstrate that using machine learning models in tandem with outcome-maximizing algorithms can improve matching decisions. While other studies have used matching, they generally ignore therapist availability as a bottleneck when assigning treatments. In contrast, we used therapist availability to estimate how well MATCH would work in practice. These results also demonstrate one economy of scale large clinics enjoy relative to smaller ones– larger practices with higher client volumes are able to leverage matching more effectively. In our case, roughly 50-200 clients per scheduling group are sufficient to yield most of the benefit of an extremely large group size. Moreover, our method could have an even greater effect in a different setting, as a recent literature review reported that therapist effects tend to be lower at universities than in other settings.[6]

Besides addressing constraints, our study differs from other matching studies that use data-based methods. The other studies have generally had samples ranging from fewer than 100 clients[4] up to a couple hundred,[2,3,5,18] except for Delgadillo et al. with 4,849 clients.[1] We had nearly double the number of observations of the largest previous study, with 8,541 clients. In addition, our data enjoy a high degree of fidelity to real-world practice. Not only were there no restrictions on either the client complaints or the therapist's choice of treatment methodology, but we also included all therapists and almost all clients, implying that the effects we observed reflect the range of real clinical outcomes and practices. Our treatment was also fairly unique in using individual therapists, as most other studies grouped therapists by personality traits or performance (in general or for a specific concern). Doorn et al. measure language-style on the individual level, but their study observed the level of matching in the clients' and therapists' language-styles and the change in clients' outcomes rather than imposing matches.[2]

The study by Delgadillo et al. heavily parallels our study, insofar as they used a tree-based ML model to do their matching using data generated by a process very similar to normal practice.[1] Our results confirm their findings that ML models can leverage the interaction of multidimensional patient features, even those not obviously connected to outcomes, to inform matching to clinicians. Our observations also agree with their finding that multiple different ML models achieve similar levels of predictive accuracy when used to estimate client outcomes from

client features and assigned therapist. Our study, however, differed in several notable ways from theirs, some of which we have already mentioned. They used groups of therapists, not individual therapists. This has the advantage of reducing the number of features but also risks losing information, as happens with any feature reduction. The interactions of the features of clients and therapists–and their effects on matching–are complex and not completely understood, so grouping therapists may obscure the therapist-level trends that otherwise matching algorithms might take advantage of. They also measured improvement using a classification framework, where the outcome was reliable and clinically significant improvement (RCSI), as measured by the PHQ-9, whereas we applied a regression framework using the drop in OQ-45 scores as the outcome. Our method thus considers gradation in outcomes more directly. Whether or not this difference in target variable alters the quality of the matches is unclear at present.

It is also unclear how our constrained matching algorithm compares to others, as no other studies address this area. Our method optimizes at two stages, first for prediction, then for total improvement for batches of clients. Queues or other approaches instead of batching could prove as effective or more effective. Moreover, directly optimizing end to end may yield a higher benefit than a two-stage optimization.

Rather than focusing on finding the best architecture for prediction, we recommend focusing on architecture related to matching. We found that the choice of ML model may not matter much in determining predictive power. The performance of the extra trees model, as measured by MAE, was only marginally better than that of the other ML models we trained (see Supplement). Rather than predictive power, the main impact of the choice of model in a matching scenario may instead be the usefulness of the resulting matches, as models with similar accuracy may still produce very different pairings of clients with therapists. For example, LASSO regression was predictive of outcomes, but the lack of interaction terms between therapist variables and client variables meant there was no actionable information for matching. Assessing which models give the greatest intervention effects under matching would require a direct way of measuring quality of matches. One possible metric is our hindsight comparison in randomly matched clients of those who met with one of their top-ten therapists versus those who did not.

Notably, we saw that clients' demographic information was also not leveraged by the ML as much as other information. When we did feature selection, biological sex, race, nationality, and

age were excluded as less relevant. In contrast, information about the client's past and concerns was retained by the model. This is a point of dissimilarity with some previous studies in which demographic was the only matching feature[24] or among the set of features.[1] Our results, however, confirms the findings outlined in Cabral and Smith[25] and in Scogin et al.[26]

As with any study, ours has limitations. We focused on measured client outcomes as given by the OQ-45 questionnaire, which is intended to measure client psychological functioning. This makes it difficult to compare to other studies about matching that used different questionnaire-based metrics, such as PHQ-9. Moreover, while using questionnaires to measure outcome improvement is convenient, the central goal of psychotherapy is not to reduce scores on a questionnaire. Indeed, in some cases (such as the treatment of trauma), immediate improvement of symptoms may not be desirable, with a more realistic trajectory being a short-term increase in symptoms followed by long-term improvement. Furthermore, different questionnaires may favor different therapeutic approaches (e.g. CBT versus interpersonal approaches). While our approach is generalizable to any numerical outcome measure (including derivatives of text- or video-based outcomes measurements such as those discussed by Kuo et al.[27]), it is important to interpret our present results in the context of symptom reduction and the OQ-45 measure specifically. Our work does not attempt to directly address the problem of how to holistically measure outcomes of psychotherapy but instead presents a matching procedure compatible with many of the proposed measures, including client dropout.

This analysis raises important questions about the potential effect of matching on the lives of individual therapists and clinic dynamics. Under our system, some therapists might meet with all the clients with a certain concern or the most stressed clients. These consequences are not clearly good or bad. On one hand, it may help therapists to specialize if they receive similar clients all the time, and it may boost their morale to receive clients fitted to their strengths; but on the other hand, some therapists might get worn out working with certain types of clients at which they excel, feel unfairly stereotyped by the algorithm (although see the supplement for an examination of algorithmic biases, which were largely null), or desire a greater variety of clients, even if their strengths as a therapist are fairly specific. The effect of client matching practices on the perceptions of clients is also unstudied, as far as the authors are aware. Our method also requires special consideration for how to add new therapists to the model.

Our work highlights the importance of considering implementation factors, like therapist capacities and scheduling details, in therapist effects studies. As far as the authors are aware, this issue has received no attention in the psychotherapy literature, although it has been addressed in an engineering study.[28] While our results indicate that a wide range of machine learning approaches give comparably accurate outcome predictions, a significant portion of the possible gain from these predictions is lost when limitations in therapist availability and limitation in client volume are considered. Success at navigating these logistical factors seems, in light of these results, to be more important than proposing new machine learning architectures, suggesting the importance of studies focused on this issue. For example, we should determine when, if ever, it makes sense to prolong wait times for the sake of seeing a particular therapist. Our results also demonstrate that substantial practical problems, such as limitations on the number of people a therapist can meet with, can be addressed while retaining much of the theoretical benefit of matching. Future study focusing on addressing other practical concerns while preserving matching benefits may likewise prove fruitful.

Methods

Data

We used clinical data for exactly 10,100 courses of individual therapy from CAPS between 2014 and 2019, with 8,541 clients. The clients consisted of 63% females and 15% of a non-white race. Client features included basic demographics, the CCAPS-62 intake survey, and OQ-45 scores collected both at intake and before subsequent appointments throughout the course of therapy. CCAPS-62 is a 62-item questionnaire developed specifically for college counseling centers that quantifies clients' psychological distress in each of eight subscales.[29] Similarly, OQ-45 is a heavily researched measure of a wide range of distress symptoms.[30] In the raw data, 93% of individual courses of therapy had an associated CCAPS intake survey and 49% had more than two OQ-45 questionnaires. The low proportion of OQ-45 scores can be explained by a particular type of therapy at CAPS, called crisis, that includes only one visit with a therapist.

In the past, clients have been matched to their therapists based on most immediate availability. Although clients occasionally had a specific therapist preference, (e.g., requested a therapist of the same gender), this was rare. Consequently, the matching between clients and therapists was

virtually random. This assumption is necessary for us to estimate the causal effect of the client-therapist match, isolating it from confounding variables.

Data cleaning and course of therapy operationalization

While we had appointment data starting in 2012, we excluded the courses of therapy prior to 2014 because we lacked OQ-45 and CCAPS data from that period. We further restricted which survey questions we considered based on qualitative criteria or incomplete data. Certain omitted questions included the clients' country of origin, academic status "other" (i.e. not a freshmen, sophomore, junior, senior, or graduate student), housing situation, area of study, and follow-up questions on race and sexual orientation (although categorical questions about race and sexual orientation were included). The full list of features can be found in the supplemental materials. We then partitioned appointment data for each client into courses of therapy, defined as a sequence of appointments, or a single appointment, with the same therapist with no more than a 180-day gap between appointments. If more than 180 days elapsed after an appointment, any subsequent appointments were considered a new course of therapy. Sometimes, a client returning to therapy after more than 180 days would not retake the CCAPS survey. Out of necessity, we excluded such courses due to lack of data. Consult the supplement for a full description of our exclusion criteria.

During the cleaning, we set aside 20% of the clients, uniformly at random, for a held-out set. The role of the held-out set is to have data that is completely independent of the training and parameter tuning processes, in order to create valid estimates of model effectiveness. The data set that we trained and tested on contained a total of 8,115 courses of therapy and 6,851 clients. The held-out data had 1,985 courses and 1,690 clients. The demographics for these data sets are only slightly different from the whole: 62.5% (train/test) and 63.3% (held-out) were female, while racial minorities account for 15.3% and 15.4% respectively. The train/test/held-out split was performed client-wise (rather than course-wise) to ensure statistical independence of the three sets.

Imputation

Clients consistently skipped particular questions on the intake survey: questions about sleep habits (92% skipped), academic concerns (92%), social life (92%), emotional well-being (92%), sexual orientation (77%), military stress (21%), pornography concern (7%), and being a ROTC

member (4%). We assumed that clients did not respond because they did not think the questions applied to them or did not want to answer. As such, we imputed values for these responses with the choice corresponding to "no" or the least concern. Some of the other questions also had missing responses but much more infrequently. For these, we applied principal components analysis (n=10), a technique that uses the variation in the data, to fill in missing values.[19]

Features and Outcome Variable

While cleaning the data, we excluded 35 features (see Supplement for a list) that were found to aid little in prediction in previous studies conducted at CAPS, ending with 131 features, including demographics (e.g. "International Student"), personal history (e.g. whether the client had ever taken medication for mental health purposes), and concern information (e.g. "There are many things I am afraid of."), as well as 113 therapists, each of which we encoded as their own binary variable. See the supplemental materials for a full list of features. We defined the outcome variable as the net reduction in OQ-45 score, which is calculated as the difference between the first OQ-45 score and the final score.

Feature selection

As a general rule, reducing the feature space from a high dimension to a lower one improves prediction modeling. We tested several feature selection methods from the Python package scikit-learn[31] using the train/test data and decided on Lasso. It gave us 42 features (counting therapists as a single feature), using cross-validation to select the penalty parameter. Table 2 has some examples of excluded features. Notably, we also tried Random Forest feature importance to select features, with similar-quality results. Before giving the features to a prediction model, we encoded therapist ID numbers into dummy binary variables.

Model selection

We employed the Python package scikit-learn to implement 30 classes of predictive models. Among these were linear SVR, lasso LARS, XGBoost,[32] gradient boosting, *k*-nearest neighbors, random forest, and extra trees (see Supplement). We trained these models, optimizing their hyperparameters, and evaluated their predicted mean absolute error (MAE), all on the same test set of data. To train the models, we used TPOT,[33] which automates hyperparameter selection and model evaluation. We emphasize that a separate held-out set was kept separate during this entire process to evaluate the final model afterwards.

Estimating optimal-therapist intervention effects

We estimated the intervention effect in two different ways. First, we estimated the effect of the predicted optimal therapist by subtracting a client's actual outcome from the predictions for each client when paired with their (predicted) optimal therapist. The difference in the means is 3.57 OQ-45 points.

Since that estimate relied on predicted values and could potentially be too optimistic, we also evaluated the therapist effect using historical observations. We determined whether each client in the held-out set was matched with one of their top ten therapists according to our model's predictions, and counted clients who did as treated with our matching. We picked ten because looking at the optimal therapist or even top-five would give only a small number of clients in the treated group, since most clients do not meet with their optimal therapist by chance. We used a two-sample t-test to compare the outcomes of the treated group from the held-out data to the outcomes in the train/test data, which acted as the controls. The difference was statistically significant with 95% confidence interval of 0.17 to 3.67. We admit, however, that exploratory analysis shows that this number might be sensitive to hyperparameter selection. In contrast to the first evaluation method, this approach is likely to be pessimistic since the top-ten therapists were used rather than the top therapist.

To find the percentiles of the distribution of client outcomes when matched with the client's optimal therapists, we injected additive noise into the predicted scores. Since our model predicts the median of a client's outcomes with that therapist, the predictions had much less variation that occurs in real life. After making some basic assumptions (detailed in Supplement), we were able to derive an estimate of the amount of noise needed. Summarizing Supplement, we estimate the noise level by calculating the model's mean squared error, subtracting off the sampling variation and subtracting the model's squared bias. To estimate the sampling variance, we trained the model on independent training sets, predicted outcomes for the same clients each time, and then calculated the variance of the predictions per client.

Matching therapists to clients

We used the Hungarian algorithm[20] to find the matches that optimize average client outcomes given (1) a group of clients, (2) a set of therapist openings, and (3) the model's predicted outcomes for each possible pairing.

Practically, using this approach corresponds to waiting until *n* slots for a course of therapy open and then assigning the next *n* clients on the waiting list to those slots using the Hungarian algorithm. A larger *n* generally yields higher client outcomes on average but could increase client wait times. In the present work, we report outcomes for a range of values of *n* between 2 and 1,000.

Estimating cohort-level matching intervention effects

We simulated therapist openings using each therapist's average caseload in our data set, and randomly sampled clients from the data to get cohorts of clients. Doing this many times and for multiple cohort sizes, we found that the average improvement as measured by OQ-45 grew rapidly as cohort size increased up to roughly 50, fully stabilizing around 500 (see Fig. 4). Of particular interest to us were the batch sizes 50 and 200, since these corresponded to the number of therapist openings in a typical week and at the beginning of a semester, respectively.

Code, model, and data availability

Complete code used to generate these results is available upon request. To protect client privacy, the trained models and data for this study can only be shared after establishing a data sharing agreement with BYU's CAPS.

Acknowledgements

The authors express thanks to BYU Counseling and Psychological Services for allowing us to use their data for this study. We also thank Saumya Sinha, Lars Nielsen, Brodrick Brown, Natalie Kirtley, Mark Beecher, Tyler Mansfield, Nathaniel Driggs, Brigham Stone Carson, Drake Brown, Matthew Gabbitas, Braeden Hintze, Eli Child, and BYU's ACME class of 2023 for helpful discussions on this work. ZMB additionally acknowledges support from the Department of Defense under contract #W911NF-18-1-0244 and the National Science Foundation under award #2137511.

Competing Interests

The authors declare no competing interests.

Author Information

Brigham Young University, Provo, UT, USA

Ammon C. Brock, Connor J. McBride, David M. Erekson, Ian H. Goodwin, Ryan K. Wood, Seth M. Peacock & Zachary M. Boyd (zachboyd@byu.edu).

Contributions

A.C.B., C.J.M, I.H.G., R.K.W., S.M.P., and Z.M.B. wrote the manuscript. A.C.B., I.H.G., R.K.W., and S.M.P developed and evaluated the algorithm. D.M.E and Z.M.B performed project management and provided conceptualizations. D.M.E performed data acquisition. All authors finalized the manuscript.

References

1. Delgadillo, J., Rubel, J. & Barkham, M. Towards personalized allocation of patients to therapists. *Journal of Consulting and Clinical Psychology* **88**, 799–808 (2020).
2. Aafjes-van Doorn, K., Porcerelli, J. & Müller-Frommeyer, L. C. Language style matching in psychotherapy: An implicit aspect of alliance. *Journal of Counseling Psychology* **67**, 509–522 (2020).
3. Boswell, J. F., Constantino, M. J., Coyne, A. E. & Kraus, D. R. For whom does a match matter most? Patient-level moderators of evidence-based patient–therapist matching. *Journal of Consulting and Clinical Psychology* **90**, 61–74 (2022).
4. Petrowski, K., Nowacki, K., Pokorny, D. & Buchheim, A. Matching the patient to the therapist: The roles of the attachment status and the helping alliance. *Journal of Nervous and Mental Disease* **199**, 839–844 (2011).
5. Werbart, A., Hägertz, M. & Borg Ölander, N. Matching Patient and Therapist Anaclitic–Introjective Personality Configurations Matters for Psychotherapy Outcomes. *Journal of Contemporary Psychotherapy* **48**, 241–251 (2018).
6. Johns, R. G., Barkham, M., Kellett, S. & Saxon, D. A systematic review of therapist effects: A critical narrative update and refinement to Baldwin and Imel's (2013) review. *Clinical Psychology Review* **67**, 78–93 (2019).
7. Crits-Christoph, P. *et al.* Meta-Analysis of Therapist Effects in Psychotherapy Outcome Studies. *Psychotherapy Research* **1**, 81–91 (1991).
8. Herman, S. M. The Relationship Between Therapist–Client Modality Similarity and Psychotherapy Outcome. *The Journal of Psychotherapy Practice and Research* **7**, 56–64 (1998).

9. Kraus, D. R., Castonguay, L., Boswell, J. F., Nordberg, S. S. & Hayes, J. A. Therapist effectiveness: implications for accountability and patient care. *Psychotherapy Research: Journal of the Society for Psychotherapy Research* **21**, 267–276 (2011).
10. Laska, K. M., Smith, T. L., Wislocki, A. P., Minami, T. & Wampold, B. E. Uniformity of evidence-based treatments in practice? Therapist effects in the delivery of cognitive processing therapy for PTSD. *Journal of Counseling Psychology* **60**, 31–41 (2013).
11. Lambert, M. J. & Bergin, A. E. *Bergin and Garfield's Handbook of Psychotherapy and Behavior Change*. (John Wiley & Sons, Incorporated, 2013).
12. Graham, S. *et al.* Artificial Intelligence for Mental Health and Mental Illnesses: An Overview. *Current psychiatry reports* **21**, 116 (2019).
13. Lutz, W. *et al.* Predicting change for individual psychotherapy clients on the basis of their nearest neighbors. *Journal of Consulting and Clinical Psychology* **73**, 904–913 (2005).
14. Aafjes-van Doorn, K., Kamsteeg, C., Bate, J. & Aafjes, M. A scoping review of machine learning in psychotherapy research. *Psychotherapy Research* **31**, 92–116 (2021).
15. Schwartz, B. *et al.* Personalized treatment selection in routine care: Integrating machine learning and statistical algorithms to recommend cognitive behavioral or psychodynamic therapy. *Psychotherapy Research: Journal of the Society for Psychotherapy Research* **31**, 33–51 (2021).
16. Idalski Carcone, A. *et al.* Developing Machine Learning Models for Behavioral Coding. *Journal of Pediatric Psychology* **44**, 289–299 (2019).
17. Bennett, K. P. & Parrado-Hernández, E. The Interplay of Optimization and Machine Learning Research. *The Journal of Machine Learning Research* **7**, 1265–1281 (2006).
18. Van Bronswijk, S. C. *et al.* Cross-trial prediction in psychotherapy: External validation of

the Personalized Advantage Index using machine learning in two Dutch randomized trials comparing CBT versus IPT for depression. *Psychotherapy Research* **31**, 78–91 (2021).

19. Stanford, D. A., Lee, J. M., Chandok, N. & McAlister, V. A queuing model to address waiting time inconsistency in solid-organ transplantation. *Operations Research for Health Care* **3**, 40–45 (2014).
20. Yang, H., Chen, R. & Kumara, S. Stable matching of customers and manufacturers for sharing economy of additive manufacturing. *Journal of Manufacturing Systems* **61**, 288–299 (2021).
21. Chang, C. C. & Lin, C. C. Dormitory Assignment Using a Genetic Algorithm. *Applied Artificial Intelligence* **35**, 2276–2297 (2021).
22. Stewart, W. J. *Probability, Markov Chains, Queues, and Simulation: The Mathematical Basis of Performance Modeling*. (Princeton University Press, 2009).
23. Beckstead, D. J. *et al.* Clinical significance of the Outcome Questionnaire (OQ-45.2). *The Behavior Analyst Today* **4**, 86–98 (2003).
24. Sterling, R. C., Gottheil, E., Weinstein, S. P. & Serota, R. Therapist/patient race and sex matching: treatment retention and 9-month follow-up outcome. *Addiction* **93**, 1043–1050 (1998).
25. Cabral, R. R. & Smith, T. B. Racial/ethnic matching of clients and therapists in mental health services: A meta-analytic review of preferences, perceptions, and outcomes. *Journal of Counseling Psychology* **58**, 537–554 (2011).
26. Bowman, D., Scogin, F., Floyd, M. & McKendree-Smith, N. Psychotherapy length of stay and outcome: A meta-analysis of the effect of therapist sex. *Psychotherapy: Theory, Research, Practice, Training* **38**, 142–148 (2001).

27. Kuo, P. B. *et al.* Machine-Learning-Based Prediction of Client Distress From Session Recordings. *Clinical Psychological Science* **0**, 0 (2023).
28. Huda, S. N. Pairing clients and psychologists using stable marriage problem approach. *IOP Conference Series: Materials Science and Engineering* **482**, 012038 (2019).
29. Locke, B. D. *et al.* Development of the Counseling Center Assessment of Psychological Symptoms-62 (CCAPS-62). *Journal of Counseling Psychology* **58**, 97–109 (2011).
30. Lambert, M. J., Gregersen, A. T. & Burlingame, G. M. The Outcome Questionnaire-45. in *The use of psychological testing for treatment planning and outcomes assessment: Instruments for adults, Volume 3, 3rd ed* 191–234 (Lawrence Erlbaum Associates Publishers, 2004).
31. Pedregosa, F. *et al.* Scikit-learn: Machine Learning in Python. *J. Mach. Learn. Res.* **12**, 2825–2830 (2011).
32. Chen, T. & Guestrin, C. XGBoost: A Scalable Tree Boosting System. in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* 785–794 (Association for Computing Machinery, 2016).
33. Olson, R. S., Bartley, N., Urbanowicz, R. J. & Moore, J. H. Evaluation of a Tree-based Pipeline Optimization Tool for Automating Data Science. in *Proceedings of the Genetic and Evolutionary Computation Conference 2016* 485–492 (Association for Computing Machinery, 2016).

Supplement

Exclusion Criteria

The appointments data had 311,168 observations. We ignored 462 appointments because there was no client ID, and 62,469 we hid away for the final held-out data set. We removed 112,506 appointments that were not for individual therapy and 21,550 more appointments because the client or therapist rescheduled. We did include no-show appointments so that we could calculate the client's attendance rate and whether the last appointment in a course of therapy was attended, though we did not use this in our analysis. From the remaining appointments, we created 20,447 courses. We dropped 3,852 because they consisted of one appointment of the type called intake. (Sometimes this appointment was with a different therapist than the rest of the therapy.) We note that one-appointment courses were not excluded as a category. We excluded 1,591 that started before the CCAPS data did, 463 that had no CCAPS survey at all, 2,883 courses of therapy that did not have surveys within the 180 days before the course started, and 3,543 that had no or just one OQ-45 score. This left us with 8,115 courses of therapy to train and test on.

Performance of other machine learning models

We trained many different models using the TPOT package (version 0.11.7) to predict the drop in OQ-45 score across a course of therapy and then tested them. TPOT automated a comparison of the performance of 30 different learning models included in its standard evaluation group, with thousands of different combinations of parameters, in order to minimize the prediction error of the drop in OQ-45 score. We closely examined the performance of seven ubiquitously used models: *k*-nearest neighbors, XGBoost, gradient boosting, Linear SVR, LassoLARS, random forest models, and extra trees. Prediction error was measured using mean absolute error (MAE).

We plotted the performance of these models to look at them more closely (see Fig. S1). For comparison, we also calculated a baseline that predicted the mean drop in OQ-45 score from the training set for every client-therapist match in the testing set. Statistically significant improvement from the baseline model suggests that the predictive model was able to learn from the data to make more accurate predictions.

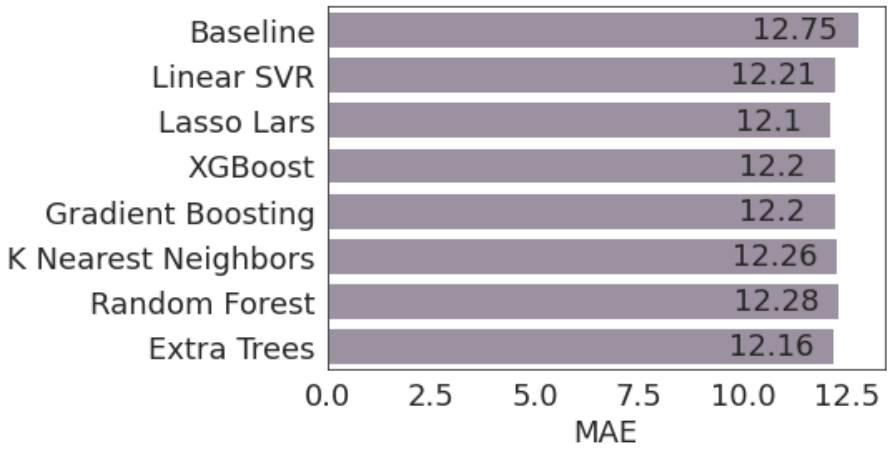

Figure S1: MAE comparison. Lasso performed marginally better than the other models, extra trees being a close second. We chose extra trees since the model already considers interaction between the client's features and the therapist. It predicted .59 OQ-45 points closer to the observed drop than the baseline. Note that the MAE in this graph was not calculated on the held-out data set.

Many models made predictions better than the baseline model. Lasso performed the best here (.65 better than baseline). We chose extra trees, however, because the Lasso model did not have interaction effects between client features and therapist and the extra trees gave similar performance. The extra trees regressor was, on average, about .59 OQ-45 points closer to predicting the true drop in OQ-45 score than the baseline model. We performed a one-sided permutation test, calculating the same metric but with the outcomes shuffled. The MAE achieved by our model was less than the MAE of each of 1000 models selected and trained by TPOT, the same way as our actual model, but using permuted data. Thus, we concluded the true prediction error of the extra trees regressor is lower than random chance ($p < 0.01$). We then used extra trees as our predictive model to assess the predicted change in OQ-45 score throughout the remainder of this study.

Biased therapist effect

Our model biases the effect of low-appointment therapists toward the mean. We analyzed this by training the model on an altered subset of our training data in which outcomes for some clients of certain therapists were artificially improved by 14 OQ points. The clients whose outcomes were artificially improved were those who had a particular concern. When trained on these data, our model predicted an average of improvement of 12.6 points compared to the unaltered data for therapists with many clients, but an improvement of just 9.34 points for therapists with few clients. In other words, the model's predictions tend to be more conservative for therapists with fewer clients, even though we know that those therapists were much more effective with some of those clients. This bias is expected, as more data points allow for greater confidence in predicting differences between two groups.

Bias on Demographics

We assessed our predictions for bias in certain demographics. First, we grouped clients together based on the quartile into which the difference between their optimal score and actual score fell, and then we ran a one-way anova test using these groups as groups and client features as the response. We used the Benjamini–Yekutieli procedure to correct for doing multiple tests. Being a racial minority and having a disability were statistically significant; the mean value for these features differed in at least one quantile from the rest. In particular, in the quartile of people whose optimal score was highest compared to their actual score, there were greater proportions of people who reported a disability and who were a racial minority than the other quartiles (see Fig. S2). These high proportions were 4% more than the lowest proportions in those quartiles that had the fewest such people. This is true in both cases. Across all our data, 9% of individuals had a disability and 15% were racial minorities. We emphasize that according to these results, benefit is higher, not lower, for these historically underserved populations than for the general population. Age also differed across the groups. The means of age differed by at most one-third of a year. It may be of interest to note that race and age were not included as features in the final feature selection and thus were not directly considered by the model.

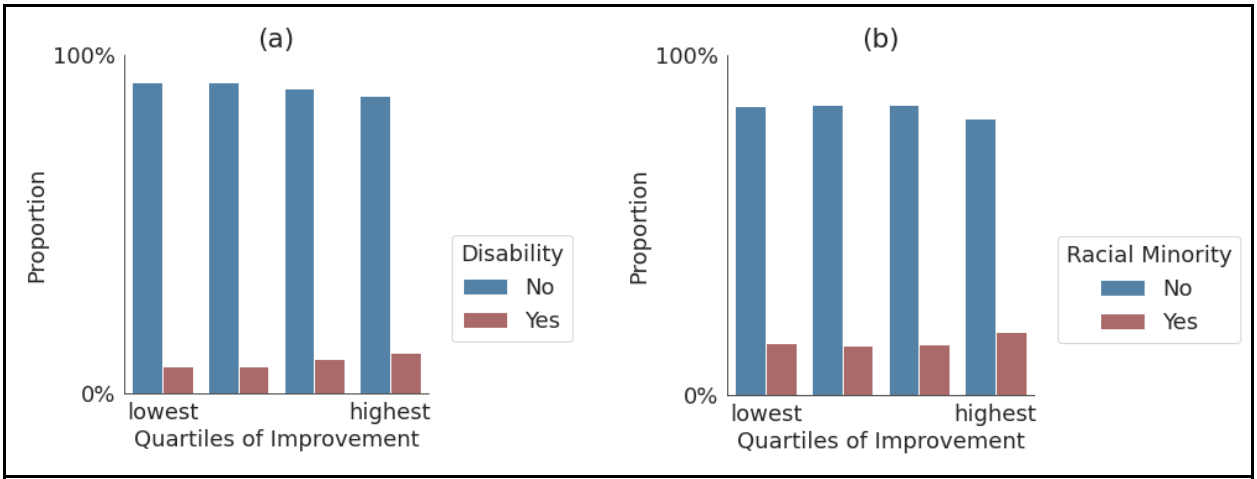

Figure S2: Grouping clients into four groups based on how much they would improve with their optimal therapist, we tested if the proportion of those of certain demographics differ across the groups. This test was significant for those with disabilities and racial minorities. We see that those who improved the most had a higher proportion of those with disabilities and racial minorities. The difference in proportion between the quartiles with the highest and lowest proportions is 4% for both features. These results are not clinically significant.

Estimators

Calculating the difference of means for two groups is a routine procedure, but not so for changes in percentiles. This required estimating the variation for each student and adjusting to the predictions so that they had the proper variance (see Fig. S3).

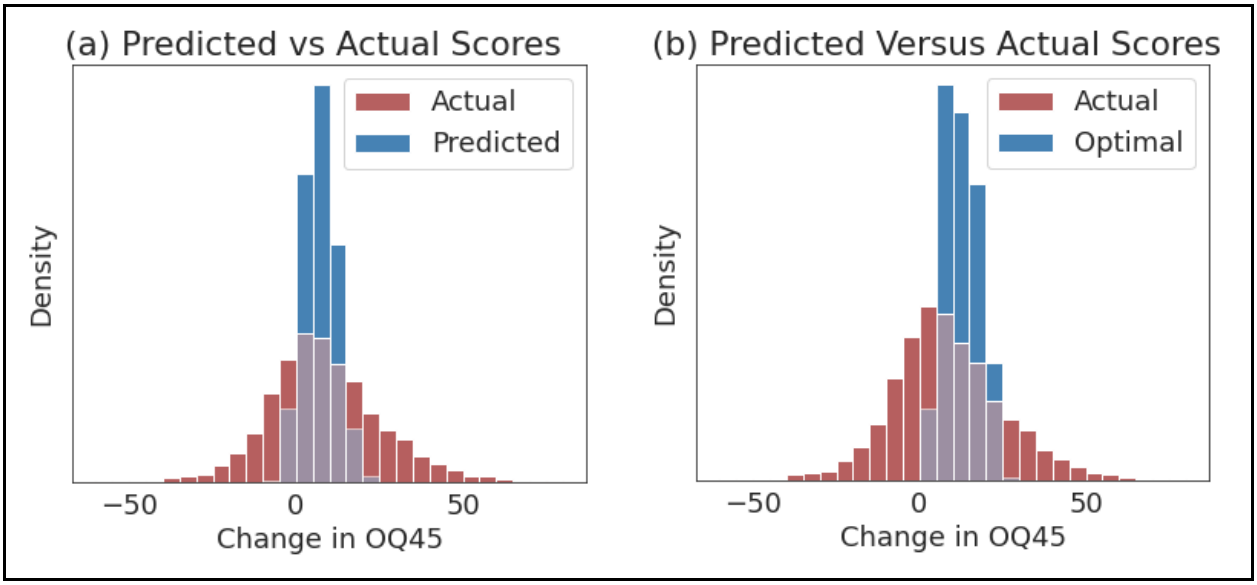

Figure S3: Spread of predicted values vs observed values.

Figure S3a: Predicted and actual. The distribution of predicted changes in OQ score has a lower standard deviation than the distribution of actual outcomes. This is because the model predicts a mean or expected value. We adjusted for this by injecting noise.

Figure S3b: Optimal and actual without noise adjustment. We predicted how the clients would do if matched with their optimal therapist. The optimal-match scores have a mean of 11.0, and actual scores have a mean of 7.4. We cannot use this graph, however, to compare percentiles.

To simplify this problem, we assumed that the only difference in each distribution of client outcomes was the mean and further assumed that each distribution was normal. Then, we defined a random variable $Y_{s,t}$, that is normally distributed with mean μ_{st} and variance σ_ϵ. The subscripted *s* and *t* remind us that the distribution depends on the student and therapist. To estimate the distribution of all outcomes, we use the predicted score as the mean and added noise, randomly sampled from a normal distribution with a mean of zero and a variance equal to our estimate of σ_ϵ. We represent a model trained on a data set *D* as a function f_D that maps a client *s* and therapist *t* to a real number. In other words, $\hat{y} = f_D(s,t)$.

The MSE overestimates σ_ϵ. It is the variance of the errors $Y_{s,t} - f_D(s,t)$. We use this formula $E[X^2] = Var(X) - (E[X])^2$ to show that MAE is a biased estimator.

$$(1) \quad E_{D\epsilon}[(Y_{s,t} - f_D(s,t))^2] = Var_{D\epsilon}(Y_{s,t} - f_D(s,t)) + (E_{D\epsilon}[Y_{s,t} - f_D(s,t)])^2$$

$$= Var_{D\epsilon}(Y_{s,t}) + Var_{D\epsilon}(f_D(s,t)) - 2Cov_{D\epsilon}(Y_{s,t}, f_D(s,t)) + (Bias_{D\epsilon}[f_D(s,t)])^2$$

Since the training data (*D)* is independent of the observed test value $Y_{s,t}$, the covariance is zero. The theoretical MSE, $E_{D\epsilon}[(Y_{s,t} - f_D(s,t))^2]$ includes not only the variation for a single client $Var_{D\epsilon}(Y_{s,t})$ but also the variance and bias of the model. Hence, the overestimation. Equation (1) implies

$$\sigma_\epsilon{}^2 = Var_{D\epsilon}(Y_{s,t}) = E_{D\epsilon}[(Y_{s,t} - f_D(s,t))^2] - Var_{D\epsilon}(f_D(s,t)) - (Bias_{D\epsilon}[f_D(s,t)])^2$$

That is, estimate the variation for a single client by subtracting the model's variance and bias from the MAE. We did this by training the model on *K* independent samples from the data and predicting outcomes for a client in the test set with each model $f_1(s,t),\ldots,f_K(s,t)$. For one client, we estimate $\sigma_\epsilon{}^2$ using the following equation,

$$\frac{1}{K}\sum_{k=1}^{K}(\,y_{s,t}-f_k(s,t))^2-\frac{1}{K-1}\sum_{k=1}^{K}\left[f_k(s,t)^2\ -\left(\frac{1}{K}\sum_{k}^{K}f_k(s,t)\right)^2\right]-\left(\frac{1}{K}\sum_{k=1}^{K}y_{s,t}-f_k(s,t)\ \right)^2$$

We take the average across all clients to get our final estimate.

Feature List

The list of features is included in the supplemental materials as a comma separated value file, named "Feature List.csv".