Applying Methods for Personalized Medicine to the Treatment of Alcohol Use Disorder

(submitted: 1/14/2020)

Abstract

Objective: Numerous behavioral treatments for alcohol use disorder (AUD) are effective, but there are substantial individual differences in treatment response. This study examines the potential use of new methods for personalized medicine to test for individual differences in the effects of cognitive behavioral therapy (CBT) versus motivational enhancement therapy (MET) and to provide predictions of which will work best for individuals with AUD. We highlight both the potential contribution and the limitations of these methods.

Method: We performed secondary analyses of abstinence among 1,144 participants with AUD participating in either outpatient or aftercare treatment who were randomized to receive either CBT or MET in Project MATCH. We first obtained predicted individual treatment effects (PITEs), as a function of 19 baseline client characteristics identified *a priori* by MATCH investigators. Then, we tested for the significance of individual differences and examined the predicted individual differences in abstinence one year following treatment. Predictive intervals were estimated for each individual to determine if they were 80% more likely to achieve abstinence in one treatment versus the other.

Results: Results indicated that individual differences in the likelihood of abstinence at one year following treatment were significant for those in the outpatient sample, but not for those in the aftercare sample. Individual predictive intervals showed that 37% had a better chance of abstinence with CBT than MET, and 16% had a better chance of abstinence with MET. Obtaining predictions for a new individual is demonstrated.

Conclusions: Personalized medicine methods, and PITE in particular, have the potential to identify individuals most likely to benefit from one versus another intervention. New personalized medicine methods play an important role in putting together differential effects due to previously identified variables into one prediction designed to be useful to clinicians and clients choosing between treatment options.

Keywords: methods for precision medicine; alcohol use disorder; Project MATCH

Public Health Significance Statements

1. This study highlights the potential use and limitations of new methods for personalized medicine with behavioral treatments for alcohol use disorder.
2. There are multiple effective behavioral treatments for alcohol use disorder, this study suggests that it may be possible to improve treatment efficacy by using algorithms to predict which treatment will work best for individual clients.

Alcohol use disorder (AUD) is characterized by considerable heterogeneity in its symptoms (Lane & Sher, 2015), as well as its clinical course (Maisto et al., 2014) and treatment effectiveness (Litten et al., 2015). One response to observed heterogeneity in treatment effects is personalized medicine, which attempts to improve overall treatment efficacy by targeting particular treatments to those individuals most likely to benefit. There are strong theoretical and empirical reasons to believe that it can be effectively applied to the treatment of AUD (Friedmann, Hendrickson, Gerstein, & Zhang, 2004; Kranzler & McKay, 2012; Litten et al., 2015; Mann et al., 2018; McKay et al., 2011; Roos, Mann, & Witkiewitz, 2017; Witkiewitz, Roos, Mann, & Kranzler, 2019).

The goal of this paper is to demonstrate the potential role that new personalized medicine methods can play in the treatment of AUD, and by extension in the treatment of other psychological disorders.

Personalized Medicine

The terms ‘personalized medicine’, ‘precision medicine’, and ‘heterogeneity in treatment effects’ are related and have been used in a variety of contexts (Kent et al., 2018; Webb et al., 2020). The premise is that individual differences are likely to exist in the effects of treatments and that if those individual differences can be predicted, they can be used to select the most effective treatment for a particular individual. In doing so, personalized medicine can improve average outcomes for the entire population (Ashley, 2015; Kranzler & McKay, 2012; M. D. Smith et al., 2013; Volkow, 2020). Much of the research in the area has involved identifying individual characteristics which interact with treatment and allow for the identification of clients who differ in treatment response (Hartwell & Kranzler, 2019). While the term ‘personalized medicine’ is relatively new, there is a long history of AUD researchers aiming to identify interactions between treatment and baseline characteristics (these must be assessed before treatment if they are be of use in helping to choose between treatment options) in predicting outcomes. The MATCH study (Project MATCH Research Group, 1998), which we use as the example in this paper, is a good demonstration of this. The study used a formal process for consulting previous research and expert opinions to make *a priori* hypotheses about differential response to treatment (Allen et al., 1997). The results of MATCH exemplify what is typically found with this type of research – support was found for only 1 of the 18 hypothesized baseline characteristics by treatment interactions. Other work in the area has thus far largely failed to find strong evidence that client data can be used guide the selection of the most effective intervention (Mann & Hermann, 2010, although see Witkiewitz, Roos, et al., 2019 for a potential exception to this). To be clear, we are not arguing against hypothesizing and testing interactions between treatment and baseline characteristics, these interactions must exist if there is any hope of realizing personalized approaches. Rather, we argue that, in many cases, heterogeneity in treatment effects is due not to a single interaction or differences between a small number of measured groups. Instead, heterogeneity is often a cumulative effect of many relatively small interactions, which are difficult to detect using standard statistical approaches.

In recent years, substantial research in the fields of statistics and machine learning has focused on using predictive methods with randomized trial data to predict treatment responses. We refer to these new methods as ‘methods for personalized medicine,’ and our primary goal is to illustrate their potential as well as their limitations for predicting outcomes in the treatment of AUD. One feature common to many of these methods is that they allow for the incorporation of many baseline covariates, which enables one to find effects that are due to many small interactions. A second feature of these methods is that they typically focus on predicting treatment response for a subgroup (Athey & Imbens, 2016; Ballarini et al., 2018; Cai et al., 2011; Seibold et al., 2018) or an individual (Ballarini et al., 2018; Lamont et al., 2016; Powers et al., 2018), and *not* on explaining the mechanisms behind these differences. That is, personalized medicine methods are designed to (ultimately) help clinicians and clients make decisions, not to help researchers understand mechanisms of behavior change or particular outcomes. The reason for this can be illustrated by the MATCH study. There were 18 hypothesized interactions in MATCH. Even with a large sample, for most effects the study was only powered to find interactions of moderate effect sizes (Allen et al., 1997) because power for interactions is typically low (McClelland & Judd, 1993). Imagine that half of the hypothesized interactions were correct, but the effect size for all of them was small, then we would have expected to find support for between 1 and 4 hypotheses (depending on the effect size), exactly what was found. Most randomized trials are not powered to find the small baseline by treatment interactions which are likely to exist. The reason personalized medicine methods work is that they combine all of these small interactions into one effect which, if it is useful, no longer has a small effect size. This enables predictions which may have clinical utility, but have limited ability to provide information about the underlying mechanisms.

Since personalized medicine methods seek to obtain predictions which are useful in clinical practice, it is important that these predictions generalize to new clients. Various analytic approaches, many drawing from machine learning are used (Athey & Imbens, 2016; Seibold et al., 2018). Additional methodological approaches can also help assure that the results are generalizable, these include: 1) identifying the baseline covariates *a priori* using theory and previous literature; 2) determining the predictive method to be used *a priori*, based on the expected differences in treatment effects; and 3) following the protocols from the original trial as much as possible. In MATCH, hypotheses about differential treatment response were made before data were collected and using existing research in the field. In this application we chose to use a logistic regression model to predict individual outcomes because predictions from this model assume two-way interactions (unless others are explicitly included) between treatment and baseline covariates, as was hypothesized in MATCH. Machine learning approaches, such as random forests (Breiman, 2001), would also capture multiway interactions and non-linear (or logit) effects. However, this comes at the price of reduced efficiency. Following the original study protocols for MATCH is more difficult as the growth curve analyses proposed are no longer considered appropriate for the primary outcomes. Additionally, it is quite difficult to obtain appropriate individual-level predictions for these count variables.

Personalized Medicine in the Treatment of Alcohol Use Disorder

Personalized medicine has the most potential to improve outcomes when a number of conditions are satisfied. First, there is evidence for heterogeneity in treatment effects. Second, different treatment options with varied mechanisms of action should exist, such that there are meaningfully different courses of treatment to choose between. Third, on average the effectiveness of a given treatment in the area is small to moderate such that there is room to improve on average treatment effects. And, fourth, there is both theoretical rationale and empirical evidence for individual differences in treatment effects. We briefly review the literature for these conditions as related to AUD.

AUD is characterized by considerable heterogeneity in its symptoms, clinical course, and treatment effectiveness (Litten et al., 2015; Witkiewitz, Litten, et al., 2019). Heterogeneity of symptoms is clear. Using the current Diagnostic and Statistical Manual for Mental Disorders, 5th edition, there are over 2000 ways to have an AUD and over 500 combinations of symptoms have been observed in population-based samples (Lane & Sher, 2015). Moreover, early attempts at classifying different types of individuals with AUD predates the scientific study of AUD (Babor, 1996), with at least 39 different attempts to classify individuals with AUD from the late 19th century to early 20th century. The clinical course of AUD is also characterized by considerable heterogeneity, with most individuals transitioning in and out of heavy drinking during the first year following treatment (Maisto et al., 2018; Witkiewitz, Maisto, et al., 2010). Treatments for AUD have been shown to be modestly effective, with effect sizes for behavioral treatments versus control conditions and for pharmacotherapies versus placebo conditions in the small to medium range (Magill & Ray, 2009; Ray et al., 2020). Thus, there is good reason to expect that personalized approaches could improve individual outcomes.

Historically there have been several behavioral approaches for treating AUD (Witkiewitz, Litten, et al., 2019). With the addition of new treatment approaches, such as combinations of behavioral, psychosocial, and pharmacological therapies, the possibilities of personalized treatment have grown (Ray et al., 2020). Disulfiram, naltrexone, and acamprosate (the three medications approved for AUD in the United States) function either as deterrents to drinking (disulfiram) or directly target neurobiological pathways associated with affective states (acamprosate), or craving and reward seeking (naltrexone). Their use has enhanced clinicians' ability to modify treatment based on client needs (Kranzler & McKay, 2012; Tomko et al., 2016; Van Der Stel, 2015). The average effect of these different treatments ranges from small to medium. Thus, for AUD there are ample treatment options, and the effect sizes for these options suggest that no one option is likely to be best for everyone.

Prior research also provides evidence that individual factors may contribute to AUD treatment outcomes. Factors previously identified include negative affective states, coping resources, comorbid psychiatric conditions and substance use, social functioning, cognitive functioning, self-efficacy, and motivation to change (Witkiewitz & Marlatt, 2004). The preliminary work for the MATCH project was notable in that it provided a set of specific hypotheses of differential response to the three behavioral treatments examined (Project MATCH Research Group, 1993). MATCH proposed a set of specific hypothesized moderating (interacting) variables, yet other research suggests that what works for a particular individual is often expected to be a function of the complex combinations of different individual characteristics and environmental factors (Babor & Del Boca, 2003; Mattson et al., 1994). A recent review of the AUD treatment literature suggests the need for consideration of numerous interacting risk factors, which previous studies identified as important, to obtain clinically useful treatment predictions (Sliedrecht et al., 2019). Such factors include not only background factors, such as comorbid psychopathology and severity of alcohol dependence, but also dynamic predictors, such as negative affective states and Alcoholics Anonymous involvement.

Recent studies provide some guidance for how personalized medicine methods may approach the field. For example, one study found that acamprosate may be most effective for individuals who primarily drink for the relieving effects of alcohol (Roos et al., 2017), whereas naltrexone may be most effective for individuals who primarily drink for the rewarding effects of alcohol (Mann et al., 2018; Witkiewitz, Roos, et al., 2019). This suggests that personalized medicine methods should include assessments of the relieving and rewarding effects of alcohol. Other work uses complex methods such as factor mixture models and network analyses to identify individual characteristics which may moderate the effects of AUD treatment (Holzhauer et al., 2017, 2020) suggesting specific baseline characteristics which should be included in predictive models and also suggesting that multiway interactions are likely. This suggests that the method used for obtaining predictions with these variables should allow for those interactions.

Ultimately, there is both theoretical and empirical evidence for specific treatment by baseline interactions in the effectiveness of treatment for AUD. If any one of those interactions was strong enough that 1 indicator (or group) was sufficient to capture the heterogeneity in response to AUD treatments, this would be our stopping point. That is, we could use that one variable to target treatment. However, what we observe for AUD appears to be that many interactions are influential in shaping response to treatment. This is where personalized medicine methods come into play. These methods are designed to bring together theory, previous research, and data to make predictions which, when validated, incorporate this information into a clinically useful tool.

Methods for Personalized Medicine

Statisticians and data scientists have responded to the interest in personalized medicine with an increased focus (Rekkas et al., 2019) on developing multiple analytic approaches to identify individuals who are likely to respond to a treatment (Basu, 2014; Cai et al., 2011; Doove et al., 2013; Foster et al., 2011; Freidlin et al., 2012; Green & Kern, 2012; Huang et al., 2012; Imai & Ratkovic, 2013; Imai & Strauss, 2011; Kent et al., 2018; Poulson, 2011; Ruberg et al., 2010; Shen et al., 2013; Zhang et al., 2013). Of the methods developed to-date, many focus on identifying subgroups of respondents who differ in treatment effects (Athey & Imbens, 2016; Ballarini et al., 2018; Cai et al., 2011; Seibold et al., 2018) A few focus on directly estimating the benefits of a treatment for an individual (Cai et al., 2011; Henderson et al., 2017; Kapelner et al., 2014; Powers et al., 2018) The focus of the current paper is on the predicted individual treatment effects approach (PITEs; Ballarini, Rosenkranz, Jaki, Konig, & Posch, 2018; Lamont et al., 2016). We believe that PITE is especially relevant to AUD treatment because of its focus on individual predictions, rather than identifying subgroups, and because it is flexible in which predictive method can be used. Many existing methods focus on identification of subgroups of individuals who respond differently to treatment. This makes sense if subgroups are expected, but with Project MATCH, a large set of continuous moderators were predicted and research on response to treatment for AUD suggests multiple moderators (Project MATCH Research Group, 1998; Witkiewitz, Hartzler, et al., 2010; Witkiewitz & Marlatt, 2004). A focus on estimating effects for individuals is appropriate when many different moderators are expected to jointly determine treatment response because this is unlikely to lead to a pattern where there are a few different groups who differ in treatment response. We note that while we think that PITE is especially compelling for AUD treatment, because most of the personalized medicine methods start from a similar premise, we expect that their results will be similar.

The basis for many personalized medicine approaches is the potential outcomes framework (Angrist et al., 1996; Holland, 1986; Rubin, 2005a) which defines individual causal effects as the difference in the potential outcome (the outcome which would have been observed) for an individual, if that individual received both treatments. Because, in most cases, only one potential outcome can be observed, the actual individual level causal effect is not observable, this is the ‘fundamental problem of causal inference’ (Holland, 1986). This framework says that if the potential outcomes can be estimated, then the effect of the intervention for a given individual can also be estimated.

The PITE approach works by using a predictive model or algorithm to estimate potential outcomes under two or more treatment conditions. Once estimates for potential outcomes are obtained, the predicted individual treatment effect is simply the difference between the potential outcomes. This difference is an estimate of the effect of one intervention versus the other (CBT vs MET in the example used in this paper) for any individual. One strength of the PITE framework is that it can be used with any predictive model or algorithm that allows individual-level outcome prediction: e.g., linear regression, multiple imputation, random forests, and Bayesian additive regression trees. The PITE approach differs from testing of interactions, which is traditionally used to understand heterogeneity in treatment effects, in that it focuses on including information from the many variables that, together, predict individual differences. Because PITE uses outcomes data along with baseline data from the original randomized trial to generate predictive algorithms, results can be used in predictions of treatment effects for people who were not part of the original RCT by using their values for the covariates with the algorithms estimated from the original trial. The PITE approach's key contributions are that it yields predictions of treatment effects for *any individual for whom covariates can be measured*, and that these predictions incorporate (at a minimum) information from all two-way interactions between treatment and the baseline covariates into one estimate.

Study Aims

There has been recognition of personalized medicine's potential to improve the efficacy of AUD treatment by using data to help determine the optimal treatments for particular individuals (Kranzler & McKay, 2012; Litten et al., 2015). Most work in this area, including the original Project MATCH trial, has examined potential moderators of treatment effects individually. In contrast, personalized medicine methods allow many predictors of individual differences to be included in one prediction, which is designed to be clinically useful. We argue that these methods are strongest when they draw on existing literature and theory to select the predictors that have theoretical or empirical support for being related to heterogeneity in treatment effects.

The goal of this study is to demonstrate the potential role that PITE (and, by extension, other personalized medicine methods) can play in the treatment of psychological disorders. We estimate individual differences in the effects of two widely used behavioral treatments for AUD: motivation enhancement therapy (MET) and cognitive-behavioral therapy (CBT). We focus on *a priori* hypothesized moderators from the original MATCH trial (Project MATCH Research Group, 1997), where it was proposed that individuals with greater alcohol involvement, cognitive impairment, comorbid psychopathology, sociopathy, alcohol-problem recognition, and motivational readiness would have better outcomes from CBT than from MET, and that those with greater anger, social functioning, network support for drinking, and self-efficacy would have better outcomes from MET than from CBT. The PITE approach provides a method for incorporating these hypothesized interactions jointly, rather than independently, into individual predictions. Specifically, we aim to use the PITE approach to demonstrate: 1) testing for individual differences in the effects of CBT and MET as a function of the originally hypothesized moderators; and 2) describing the predicted individual effects of CBT and MET including predictive intervals. We then explore both the clinical and theoretical implications of these results.

Method

The multi-site RCT Project MATCH trial recruited clients from 1991 to 1993 (Project MATCH Research Group, 1997). Although this trial is now dated, two of its features – that it was a randomized trial comparing the effects of treatments that are still widely used for AUD, and that it posited specific hypotheses about factors underlying differences in the effects of treatment – make this dataset suitable to the present investigation. Participants (n=1,726) were recruited from two populations: outpatient treatment (n=952) or aftercare following inpatient treatment (n=774). The project’s Coordinating Center was responsible for randomization of participants into three treatment groups through a probabilistic balancing procedure. These groups were CBT (n=567), MET (n=577), and 12-step facilitation (TSF; n=584). To simplify this demonstration of the PITE approach, the current study compares only the CBT and MET treatments. Treatment lasted 12 weeks, and consisted of either 12 CBT sessions delivered weekly, or 4 MET sessions delivered in the first, second, sixth, and twelfth weeks. Further details of the study design and methodology are described in the outcomes paper (Project MATCH Research Group, 1997).

Within Project MATCH, the goal of CBT (Kadden et al., 1995) was to achieve and maintain abstinence from alcohol by finding healthier ways to manage life stressors and distress. CBT therapists assumed client motivation to do so already existed, and therefore focused on identifying and changing maladaptive thinking and coping skill deficits, rather than on motivational factors (Gaston et al., 1998; Waddington, 2002). This was accomplished by the therapists’ provision of coping and drink-refusal skills, which were frequently practiced through role-playing and rehearsal. The goal of MET (W. Miller et al., 1995), in contrast, was to mobilize the person’s own commitment and motivation to change. Following Prochaska and DiClemente's (1982) stages of change model, MET therapists helped clients to examine the effects of drinking on their lives and to develop and implement plans to stop drinking (Gaston et al., 1998).

Participants

Project MATCH eligibility criteria were being at least 18 years old, self-reporting drinking within the three months prior to entering the study; having a current DSM-III-R diagnosis of alcohol abuse or dependence (American Psychiatric Association, 1987), with alcohol as the principal drug of abuse; lacking any open legal or probation/parole requirements that could impede participation; and having the ability to read at a sixth-grade level. Women accounted for 24% (n=420) of the total sample, 80% of participants identified themselves as white, the median age was 38; the youngest was 18 and the oldest, 76 (Carroll et al., 1998; Longabaugh et al., 2001).

Measures

The primary measures used in the present study included 1) the variables selected for the original MATCH predictions, and 2) individual-level demographic variables, both of which were used in predicting abstinence from alcohol at a follow-up assessment conducted 15 months after treatment began (that is, one year after treatment ended). Only the baseline values of these predictors were used for predicting treatment response as in any clinical application only baseline data would be available for making treatment selection.

Outcome measure

The outcome for the current study was a binary indicator of abstinence versus any drinking at the 15-month follow-up assessment (W. R. Miller, 1996). Given the relapse and remission patterns that occur during the first year following treatment (Maisto et al., 2018) we were interested in outcomes at the last follow-up assessment. This measure was dichotomized from the variable, percent days abstinent in the last 90 days from the original MATCH trial. Abstinence was selected instead of number of drinking days to simplify the PITE models. Because 37% of clients in the outpatient group and 56% in the aftercare group reported no drinking, this abstinence outcome captures much of the variability in drinking. Further, the procedure for using PITE with binary outcomes is established, while count outcomes require additional development.

Variables from the original MATCH predictions

Project MATCH tested a set of primary and secondary matching hypotheses with variables that had been previously researched in single-site studies. We used these matching variables to guide our inclusion of baseline covariates for the present study (Longabaugh et al., 2001), with only one exception: we did not include in our analyses one of the original matching variables, “Conceptual Level,” due to study authors’ concerns about the quality of that variable. Baseline covariates that were included are described next. The Assertion of Autonomy Scale (AAS) of the Interpersonal Dependency Instrument (IDI; Hirschfeld et al., 1977), measured the extent to which clients report being indifferent towards the opinions of others, and thus, how important it is for an individual to have the approval of his/her loved ones. The Psychosocial Functioning Inventory (PFI; Feragne, Longabaugh, & Stevenson, 1983) was designed to measure self-reported psychosocial functioning and overall well-being. The Social Supports-Friends & Family (SS1 & SS2; Sarason, Levine, Basham, & Sarason, 1983) measured individuals’ perceptions of the social support they received from friends and family, and their satisfaction with such support. The Alcohol Abstinence Self-Efficacy Scale (AASE; DiClemente, 1986; DiClemente, Carbonari, Montgomery, & Hughes, 1994) assessed respondents’ self-efficacy and confidence about abstaining from drinking when faced with temptations using two separate dimensional scales. The University of Rhode Island Change Assessment (URICA; DiClemente & Hughes, 1990) consists of five scales, including – pre-contemplation, contemplation, determination, action, and maintenance – were used to measure readiness for change. Cognitive impairment was measured using the Cognitive Impairment Index, a composite measure resulting from summing standardized scores from a number of cognitive assessments (Shipley Institute of Living Scale (Shipley, 1940), the Trail Making Test (Reitan, 1958), and the Symbol Digit Modalities Test (A. Smith, 1973)). Higher values on this index indicated higher levels of impairment.

We further included the Addiction Severity Index (ASI; McLellan, Luborsky, Woody, & O'Brien, 1980), the Alcohol Anonymous Involvement scale (AAI, Tonigan, Connors, & Miller, 1996), a composite measure of drinking severity and negative consequences of drinking (Magura et al., 2013), the Religious Beliefs and Background (RBB), the Seeking of Noetic Goals (SONG; Crumbaugh, 1977), alcoholism typology (Brown et al., 1994), and the State-Trait Anger Scale (TAS Form-90; Miller, 1996). Participants' sex, age, number of drinks per drinking day and percent days abstinent were also included as they are often examined as baseline covariates in studies of AUD treatment outcomes.

Data Analysis

While individual variables in the analyses had no more than moderate amounts of missing data (0.0% to 9.1%), listwise deletion would have resulted in large amounts of missingness (42.0%). Therefore, we performed single imputation for the CBT and MET treatment conditions separately by condition to allow for analysis of interactions. Single imputation results in unbiased parameter estimates under the assumption that data are missing at random, although it will result in *p*-values from inferential statistics being too low (Schafer, 1999). The only inferential tests used in this paper are the permutation test and individual predictive intervals described below. We used single imputation for this demonstration of the PITE approach because there is not yet any approach for adjusting these tests for multiple imputation. As a result, we acknowledge that the estimation of *p*-values and predictive intervals will be somewhat liberal and as such, should be interpreted with caution (Schafer, 1999). Imputation used predictive mean matching for continuous variables, logistic regression imputation for binary variables, polytomous regression for nominal categorical variables, and a proportional odds model for ordered categorical variables with more than 2 categories. The imputation model included 109 baseline covariates, and also drinking outcomes (including number of drinks per drinking day, percent days abstinent, and abstinence) at one year following treatment.

The full Project MATCH sample included some individuals who were recruited from outpatient treatment, and others who were part of an aftercare sample who had recently received inpatient treatment. Due to the substantive differences between these two patient populations as judged by the original Project MATCH researchers (Project MATCH Group, 1997), we generated PITE predictions and performed permutation tests on each population separately. Predicted individual treatment effects were obtained using a previously published procedure (Lamont et al., 2016) which we describe below.

PITE directly applies a potential outcome framework: the predicted individual treatment effect (PITE) for each client *i*, is estimated as the difference in the predicted (or potential) outcome Y*, given observed covariates X and the predictive method used, under experimental treatment (T = 1) and the predicted value under control condition or an alternative treatment (T = 0):

$$\widehat{PITE}_i = \left(Y_i^{¿} \vee X = X_i, T = 1\right) - \left(Y_i^{¿} \vee X = X_i, T = 0\right) \qquad \text{Eq. 1}$$

Where $\widehat{PITE}_i$ is the predicted treatment effect for individual i, and X_i is a vector of the baseline covariates for individual i. $Y_i^{¿}$ are the predictions obtained from the predictive models using the individual's observed values on the covariates.

We describe the computation of PITE for the MATCH data in four steps:

1) Fit a logistic regression model (chosen because it is congruent with the MATCH hypotheses) for clients who were randomized to receive CBT in which we predict their outcome using the a priori identified baseline covariates plus demographics.
2) Fit a logistic regression model for clients who were randomized to receive MET in which we predict their outcome using the same variables in step 1.
3) Using the results from steps 1 and 2 above, compute the predicted probability of abstinence under MET and CBT for *every* individual *i* using their baseline covariates.
4) From Equation 1, for each individual client subtract their probability of abstinence under CBT from their probability of abstinence under MET. This difference is the PITE.

PITE values of 0 indicate that the two treatments are expected to have the same effect for a client, positive values mean that the client is more likely to be abstinent under MET, and negative values mean that the client is more likely to be abstinent under CBT.

To obtain PITE, predictive models for clients in the aftercare and outpatient populations, 19 baseline covariates, consisting of a combination of MATCH variables identified *a priori* and baseline demographics were used as predictors of abstinence at the 15-month mark for those patients who had received CBT (step 1) and for those who had received MET (step 2).

One of the first questions to ask when conducting these analyses is whether the individual differences observed in the PITEs are greater than those that would arise by chance. The individual differences in the PITEs were quantified by the observed standard deviation (SD) of the PITEs across all individuals in the sample. The SD of the PITEs is, by definition, the average differences between individuals in the predicted treatment effects. If the SD of the PITEs were 0 it would mean everyone has the same predicted treatment effect. We then estimated the sampling distribution of this standard deviation (SD) under the null hypothesis that differences observed between individuals were due to chance using a permutation test (Chang et al., 2019). For this test, 1000 bootstrap samples were drawn from the data and treatment condition was randomly permuted such that for each bootstrap sample differences in predictions between treatment conditions were only due to chance (average treatment effects were removed). PITEs were then calculated for each bootstrap sample, the SD of PITE from the sample was computed, and the sampling distribution of the SD of the PITEs is calculated as the distribution of the SDs across all bootstraps. The *p*-value for the permutation test is the proportion of the permutations where the bootstrap SDs are greater than the PITE SD from the MATCH dataset given the actual treatment condition (Rosenbaum, 1984; Rubin, 2005b).

One method to quantify how large the differences in predicted treatment effects are across individuals is to examine the distribution of the PITEs. Here we report on the person at the 25th percentile and the person at the 75th percentile as half the sample falls between these individuals with the other half being more extreme.

The individual-level PITEs are much more useful if we are also able to provide predictive intervals for each individual. For a given client, while it is useful to know that they are more likely to be abstinent under CBT than MET, it is also useful to know that 80% of the time this client is expected to do better under CBT. To obtain these individual-level predictive intervals, we took the model parameter estimates and standard errors for the response under treatment and separately under control. This allows us to compute the individual-level sampling distribution for the probability of abstinence under treatment and under control for each individual, given their observed values of the predictors. We then randomly drew 100,000 times from each distribution, computed the difference of the two draws, and across all 100,000 draws we then computes individual-level predictive intervals. We note that, across individuals, these intervals can be substantially different depending on which variables contribute to that different individual's predictions which impacts the reliability of the estimated effects for each individual. In this paper we report the 20th and 80th percentiles for these intervals. While different percentiles can be chosen, we argue that these provide a good balance for selecting an intervention for an individual client. Smaller intervals (40th and 60th percentiles) would not be much better than chance. Wider intervals (5th and 95th percentiles) would provide greater certainty that the intervention would benefit the client, but at the cost of having few clients for whom the PITE clearly suggests choosing one treatment over the other. If the interventions had different costs and/or side effects, then other intervals which favored the cheaper or less risky intervention, unless there is clear evidence for the alternative, would be preferred.

Results

We started our analyses by examining the distributions of the *a priori* MATCH covariates, demographic variables, and outcomes for both the aftercare and outpatient samples at baseline to validate the consensus among researchers familiar with the MATCH trial (Connors et al., 1996; Maisto et al., 2015) that these should be considered separate populations. Table 1 shows means and SDs for each of the baseline covariates included in the PITE predictions, as well as prevalence of abstinence, by aftercare and outpatient samples. The same table also presents the results of *t*- and chi-square tests that were conducted to assess the differences between the outpatient and aftercare samples. The groups of participants represented by each condition were found to differ significantly on 14 of the 19 measures examined, providing further support for separating these two groups because they represent substantively different populations of those in treatment for AUD. We ran further *t*- and chi-square tests to evaluate differences in key variables between treatment conditions. The only significant differences found were in the outpatient sample on alcoholism typology and in the aftercare sample on Alcoholics Anonymous involvement and severity of drinking consequences.

Individual differences in CBT versus MET among the aftercare sample

PITE analyses start by answering the global question: is there evidence of significant individual differences in the effects of CBT versus MET? The results pertain only to the hypothesized baseline covariates and predictive method used. They would be somewhat different if other covariates were included or a different predictive method was employed. The SD of the PITEs for the 527 individuals in the aftercare sample who had received either CBT or MET was 0.22. This means that the predicted probability of abstinence for the average individual was 0.22 from the average treatment effect. The permutation test showed that the mean of the sampling distribution for the SD of the PITEs, under the null hypothesis of homogeneity in treatment effects, was 0.18, and the *p*-value for the significance of the observed SD from the permutation test was greater than .05. Thus, there is not significantly more heterogeneity in the effects of these treatments than would be expected due to chance. In practical terms, individuals in the aftercare sample did not respond significantly differently to the two treatment conditions on the basis of the predictions from the logistic model and baseline covariates used.

Individual differences in CBT versus MET among the aftercare sample

In the outpatient sample of 617 MATCH participants who received either CBT or MET the observed SD of the PITE for was 0.23 and the mean SD across permutations (given only chance heterogeneity in treatment) was 0.17. The permutation test for the outpatient sample found that 24 of 1,000 permutations resulted in SDs greater than that observed in the data, thus the p-value for the permutation test was .024. Practically speaking, this means that, together, the

19 covariates significantly predicted individual differences in the effects of CBT versus MET in the outpatient sample.

Quantifying individual differences in the effects of CBT and MET in the outpatient sample

Since we found significant individual heterogeneity in treatment effects in the outpatient sample, we proceed by exploring those individual differences, Figure 1 presents the PITE predictions (circles) and predictive intervals for a random sample of 100 individuals. In theory the distribution of the PITEs could range from -1 (the client is predicted to be abstinent with 100% certainty under CBT and to be non-abstinent with 100% certainty under MET) to 1 (the opposite effect is predicted). In practice, observed PITEs ranged from -0.75 to 0.56, with a median predicted value of -0.08 (meaning that an individual at the 50th percentile is expected to have an 8% increase in the probability of abstinence if given CBT versus MET); and 62.7% of the sample had a PITE score of less than 0, indicating that they would be expected to do better with CBT. To quantify the practical impact of the heterogeneity in effects observed, we examine an individual at the 25th percentile (see Figure 1). This individual happens to be a 28-year-old white male with 11 years of education, his PITE was -0.24 indicating that he is predicted to be 24% more likely to be abstinent under CBT than MET. We compare this client to the client at the 75th percentile who is also a white male, but is 38 years old and has an education of 18 years. This client has a PITE of .09 indicating that he is 9% more likely to be abstinent under MET than CBT. The PITEs for half of the sample fall between these two individuals and half of the sample is more extreme with an expected change in the probability of abstinence of over 33% when selecting one treatment over the other.

While the results from the permutation test show that the differences in the PITEs in Figure 1 are significant, the focus of PITE is the utility of these individual predictions. For a client and their clinician, it is useful to have not just the prediction but to have some measure of the reliability of the prediction. Thus, Figure 1 also includes an interval for each person bounded by the 20th percentile on the bottom and the 80th percentile on the top. We argue for these intervals over the traditional 95% interval because they are for individual predictions rather than population estimates. Although this interval may seem small from a statistical perspective, this interval tells us that, for a specific client, 4 out of 5 times, that client would do better with a particular treatment. We argue this is a useful metric for making a clinical decision, although any interval could be used, and the interval can be easily modified when one treatment costs more, takes longer, or has more side effects then the other. As can be seen in Figure 1, for many individual clients the predictive intervals include 0. This means that based on their data we are not confident that these clients would respond differently to the two treatments. Across all 617 individuals in the outpatient data, we found 228 (37%) whose entire interval was less than zero, and thus, were predicted to do better with CBT at least 80% of the time. There were 98 clients (16%) whose entire interval was greater than zero, and thus, were predicted to do better with MET at least 80% of the time. If this decision rule were used, PITEs would have resulted in a preference to assign specific clients, based on individual characteristics, to either CBT (37% of clients) or MET (16% of clients) for 53% of the outpatient clients in MATCH whereas we would have less confidence about the utility of the predictions for the other 47%.

Obtaining PITE for a New Client

This method can also be easily applied to clients who did not participate in the original study. To demonstrate, we randomly selected an individual from the twelve-step condition which was not included in this analysis, and extracted their values on the 19 baseline covariates. This randomly chosen individual was a 62 years old, white male. He drank heavily at baseline and had experienced severe consequences as a result of his drinking. On days that he drank, he had an average 4 drinks a day and he drank 96% of days. He had high levels of support from friends, but lower than average support from his family.

Based on these baseline values, and data on each of the other variables listed in Table 2, we calculated his PITE and predictive interval using the algorithm derived from the logistic regression model estimates presented in Table 2. His PITE was -0.16, which indicates that he was predicted to be more likely to remain abstinent under CBT. However, his 80% predictive interval (-0.377, 0.096) includes 0, indicating that there is a reasonable chance that his outcome would be the same under either CBT or MET.

This same algorithm can be used for any client for whom data can be collected on the 19 variables included in the PITE predictions. Collection of this information might be most easily implemented using computer-assisted self-interviewing techniques. The survey would replicate the items for each of the measures included in the PITE calculation. These responses would be inserted into the logistic regression equation provided in Table 2 and the algorithm described above would be used to obtain the individual’s PITE.

Comparison of PITE results to independent tests of MATCH hypotheses

Because our outcome and model was different from that used in testing the MATCH hypotheses (Allen et al., 1997), we ran analyses in which we tested whether each of the 19 variables included in the PITE estimation was a moderator of the treatment effects. All analyses included only baseline drinking (percent days drinking) as a covariate, the covariate being tested for moderation, treatment, and the interaction between treatment and the moderator. Results showed that for the aftercare sample none of the 19 interactions, run separately, were significant in predicting differences between CBT and MET. For the outpatient sample 3 of the 19 interactions were significant (the moderators identified were IDI, AASE-C, and AASE-T). From a binomial distribution there is a 7% chance of finding 3 or more of 19 interactions to be significant (with type I error set to .05), if, in reality, there are no interactions between treatment and control. This result is ambiguous, one interpretation could be that some of the moderation hypotheses were true, but the effect sizes were small.

Discussion

In the field of AUD treatment, much previous research has gone into matching clients to the treatment which will be most effective for them. The present study illustrates how new methods can build on these previous results to create clinically useful predictions of the treatment effect expected for an individual client. More than 30 years ago, when Project MATCH was first initiated, researchers were interested in testing specific hypotheses predicting which of the most commonly-used AUD treatments would be more effective for certain individuals. That project's results largely failed to find support for these *a priori* hypotheses about differential effectiveness. Using the PITE framework, the present study found evidence of significant and meaningful heterogeneity in the effects of CBT and MET for those in the Project MATCH outpatient (but not aftercare) sample. This finding illustrates both the potential and the limitations of these methods. Personalized medicine methods in general, and PITE in particular, are well suited for testing global hypotheses about individual differences because of their ability to combine many small effects into one estimate of heterogeneity. However, we argue that this very feature of these approaches means that they are not well-suited for exploring or confirming hypotheses about the causes of that heterogeneity. For example, a particular variable could be important because it predicts treatment compliance, another variable could matter for only a subgroup of respondents, and other variables could independently have only small interactions with treatment. While methods, such as measures of variable importance (Bagherzadeh-Khiabani et al., 2016; Strobl et al., 2008) do exist for giving applied researchers some ideas about what is underlying the results, we argue that those should be seen as, at best, exploratory, and not tests of existing hypotheses such as were made in the MATCH project.

We contend that the intention and greatest value of personalized medicine methods is not in testing theory or establishing that different mechanisms of change exist, but in providing clinically useful predictions for individual clients and their clinicians. Just over 50% of the MATCH participants in the outpatient sample were predicted, with at least 80% certainty, to be more likely to be abstinent one year after treatment with CBT (37%) or MET (16%). This is in contrast to the take home message from MATCH, which is that the different behavioral therapies performed about the same with minimal evidence for individual differences (Project MATCH Research Group, 1998). Further, the difference in effect sizes predicted under CBT and MET were not trivial; with the difference between someone at the 25th and someone at the 75th percentile in the predictions being a 33% change in the probability of abstinence. We additionally demonstrated how predictions and intervals can be obtained for a new patient not in the sample. PITEs can be used to obtain predictions for any new client with the only requirement being that the covariates used be assessed.

While different methods have been proposed for personalized medicine, we argue that the PITE approach is particularly useful in this situation because: 1) it provides specific individual predictions, something which few other methods do; 2) it provides an overall test for whether significant heterogeneity exists, also a feature which has not been widely implemented in other methods; and 3) it can provide individual-level predictive intervals. This last feature of PITE is somewhat straightforward for parametric statistical approaches but is much more difficult, and to our knowledge, has not yet been implemented with machine learning predictions. A strength of the PITE approach is that it puts information about expected outcomes and uncertainty in the hands of clinicians and clients so that this information may be used in treatment decisions. This is in contrast to personalized medicine approaches which simply use data to say which treatment is expected to be best for the client. While we do provide the equations in Table 2 for obtaining individual level predictions from the 19 variables used here, we do not believe that the conditions are currently in place for using these predictions in clinical practice. In our view before any personalized medicine predictions are implemented it is critical that the predictions be replicated in independent data. This approach would involve costs (collecting and inputting data) and time on the part of clients and clinicians, and we want to see proof that these methods work as expected in new clients before implementation. In the case of MATCH data, it is also important to note that new treatments are now available and new research on heterogeneity in treatment effects has been published. Updated results that include pharmacological and combined interventions are needed.

One strength of the current study is that, in contrast to many existing applications of personalized medicine, the MATCH trial had specific *a priori* hypotheses about which factors were likely to drive treatment response. While PITE and other methods for examining personalized medicine can be utilized in a completely data-driven manner, we argue that the results are more convincing when the selection of the variables for predicting heterogeneity in treatment response are based on specific hypotheses. Thus, this proof-of-concept study shows how personalized medicine methods may be useful in the treatment of AUD. However, it also has various limitations. These include that the patients recruited for MATCH exhibited relatively little psychiatric or substance comorbidity, and were treated in the context of a research study.

The simplification of abstinence as the only outcome is also a limitation, given that drinking reductions short of abstinence are also desirable (Witkiewitz, Wilson, et al., 2019). The age of the MATCH dataset and the number of prior analyses conducted using it should also be taken into account in the context of limitations.

It is also important to note that more methodological work is needed for these methods of personalized medicine. A few of the areas where more attention is needed include: 1) there is not yet any method for correcting the *p*-value of the permutation test for missing data; 2) methods need to be developed to compare and choose between predictive approaches; 3) methods are needed for comparing different personalized medicine methods; 4) predictive intervals are needed for any viable prediction method as well as for approaches to personalized medicine other than PITE; and, 5) research will be needed on how these methods can be effectively and efficiently implemented in practice. Research into best practices around all of these decisions is only now beginning and much more work is needed.

In sum, while this study's results may not yet provide a basis for guiding clinical decisions, they do open the door to the possibility that methods from personalized medicine, including but not limited to the PITE approach, could improve clients' outcomes by providing guidance for their choice of treatment options. The original MATCH study was the largest study of its time to attempt to test this proposition, and the current results provide support for some of its original *a priori* matching hypotheses while illustrating that it continues to be difficult to know which of these specific hypotheses is true. A major point of this paper is that from the perspective of personalized medicine, what ultimately matters is not the specific mechanisms but whether clinically useful predictions can be made.

References

Allen, J., Anton, R. F., Babor, T. F., Carbonari, J., Carroll, K. M., Connors, G. J., Cooney, N. L., Del Boca, F. K., DiClemente, C. C., Donovan, D., Kadden, R. M., Litt, M., Longabaugh, R., Mattson, M., Miller, W. R., Randall, C. L., Rounsaville, B. J., Rychtarik, R. G., Stout, R. L., … Zweben, A. (1997). Project MATCH secondary a priori hypotheses. *Addiction*, *92*(12), 1671–1698. https://doi.org/10.1111/j.1360-0443.1997.tb02889.x

American Psychiatric Association. (1987). *Diagnostic and Statistical Manual of Mental Disorders* (3rd ed., r). Author.

Angrist, J. D., Imbens, G. W., & Rubin, D. B. (1996). Identification of Causal Effects Using Instrumental Variables. *Journal of the American Statistical Association*, *91*(434), 444–455. https://doi.org/10.1080/01621459.1996.10476902

Ashley, E. A. (2015). The precision medicine initiative: A new national effort. *JAMA - Journal of the American Medical Association*, *313*(21), 2119–2120. https://doi.org/10.1001/jama.2015.3595

Athey, S., & Imbens, G. (2016). Recursive partitioning for heterogeneous causal effects. *Proceedings of the Natonal Academy of Science of the United States of America*, *113*(27), 7353–7360. https://doi.org/10.1073/pnas.1510489113

Babor, T. F. (1996). The classification of alcoholics: Typology theories from the 19th century to the present. *Alcohol Health and Research World*, *20*(1), 6–14.

Babor, T. F., & Del Boca, F. K. (Eds.). (2003). *Treatment Matching in Alcoholism*. Cambridge University Press.

Bagherzadeh-Khiabani, F., Ramezankhani, A., Azizi, F., Hadaegh, F., Steyerberg, E. W., &

Khalili, D. (2016). A tutorial on variable selection for clinical prediction models: feature selection methods in data mining could improve the results. *Journal of Clinical Epidemiology*, *71*, 76–85. https://doi.org/10.1016/j.jclinepi.2015.10.002

Ballarini, N. M., Rosenkranz, G. K., Jaki, T., Konig, F., & Posch, M. (2018). Subgroup identification in clinical trials via the predicted individual treatment effect. *PLOS One*, *13*(10), 1–22. https://doi.org/10.1111/biom.12522

Basu, A. (2014). Estimating person-centered treatment (PeT) effects using instrumental variables: An application to evaluating prostate cancer treatments. *Journal of Applied Econometrics*, *29*, 671–691. https://doi.org/10.1002/jae.2343

Breiman, L. (2001). Random Forests. *Machine Learning*, *45*, 5–32. https://doi.org/10.1023/A:1010933404324

Brown, J., Babor, T. F., Litt, M. D., & Kranzler, H. R. (1994). The Type A/Type B distinction: Subtyping alcoholics according to indicators of vulnerability and severity. In T. F. Babor, V. M. Hesselbrock, R. E. Meyer, & W. Shoemaker (Eds.), *Annals of the New York Academy of Sciences, Vol. 708 Types of alcoholics: Evidence from clinical, experimental, and genetic research.* (pp. 23–33). New York Academy of Sciences.

Cai, T., Tian, L., Wong, P. H., & Wei, L. J. (2011). Analysis of randomized comparative clinical trial data for personalized treatment selections. *Biostatistics*, *12*(2), 270–282. https://doi.org/10.1093/biostatistics/kxq060

Carroll, K., Cooney, N., Donovan, D., Longabaugh, R., Wirtz, P., Connors, G., DiClemente, C., Kadden, R., Rounsaville, B., & Zweben, A. (1998). Internal validity of project match treatments: Discriminability and integrity. *Journal of Consulting and Clinical Psychology*, *66*(2), 290–303. https://doi.org/10.1037/0022-006X.66.2.290

Chang, C., Jaki, T., Sadiq, M. S., Kuhlemeier, A. A., Feaster, D., Cole, N., Lamont, A., Oberski, D., Desai, Y., & Van Horn, M. L. (2019). *A Permutation Test for Assessing the Presence of Individual Differences in Treatment Effects*. http://arxiv.org/abs/1911.07248

Connors, G. J., Maisto, S. A., & Donovan, D. M. (1996). Conceptualizations of relapse: A summary of psychological and psychobiological models. *Addiction*, *91*(12), S5–S13. https://doi.org/10.1111/j.1360-0443.1996.tb02323.x

DiClemente, C. C. (1986). Self-Efficacy and the Addictive Behaviors. *Journal of Social and Clinical Psychology*, *4*(3), 302–315.

DiClemente, C. C., Carbonari, J. P., Montgomery, R. P. G., & Hughes, S. O. (1994). The alcohol abstinence self-efficacy scale. *Journal of Studies on Alcohol*, *55*(2), 141–148. https://doi.org/10.15288/jsa.1994.55.141

DiClemente, C. C., & Hughes, S. O. (1990). Stages of change profiles in outpatient alcoholism treatment. *Journal of Substance Abuse*, *2*(2), 217–235.

Doove, L. L., Dusseldorp, E., Van Deun, K., & Van Mechelen, I. (2013). A comparison of five recursive partitioning methods to find person subgroups involved in meaningful treatment–subgroup interactions. *Advances in Data Analysis and Classification*, *8*(4), 403–425. https://doi.org/10.1007/s11634-013-0159-x

Feragne, M. A., Longabaugh, R., & Stevenson, J. F. (1983). The Psychosocial Functioning Inventory. *Evaluation and the Health Professions*, *6*(1), 25–48.

Foster, J. C., Taylor, J. M. G., & Ruberg, S. J. (2011). Subgroup identification from randomized clinical trial data. *Statistics in Medicine*, *30*(24), 2867–2880. https://doi.org/10.1002/sim.4322

Freidlin, B., McShane, L. M., Polley, M.-Y. C., & Korn, E. L. (2012). Randomized phase II trial designs with biomarkers. *Journal of Clinical Oncology*, *30*(26), 3304–3309. https://doi.org/10.1200/JCO.2012.43.3946

Friedmann, P. D., Hendrickson, J. C., Gerstein, D. R., & Zhang, Z. (2004). The effect of matching comprehensive services to patients' needs on drug use improvement in addiction treatment. *Addiction*, *99*(8), 962–972. https://doi.org/10.1111/j.1360-0443.2004.00772.x

Gaston, L., Thompson, L., Gallagher, D., Coumoyer, L., & Gagnon, R. (1998). Alliance, technique, and their interactions in predicting outcome of behavioral, cognitive, and brief dynamic therapy. *Psychotherapy Research*, *8*(2), 190–209. https://doi.org/10.1080/10503309812331332307

Green, D. P., & Kern, H. L. (2012). Modeling heterogeneous treatment effects in survey experiments with bayesian additive regression trees. *Public Opinion Quarterly*, *76*(3), 491–511. https://doi.org/10.1093/poq/nfs036

Hartwell, E. E., & Kranzler, H. R. (2019). Pharmacogenetics of Alcohol Use Disorder Treatments: An Update. *Expert Opinion on Drug Metabolism and Toxicology*, *15*(7). https://doi.org/https://doi.org/10.1080/17425255.2019.1628218

Henderson, N. C., Louis, T. A., & Rosner, G. L. (2017). *Individualized Treatment Effects with Censored Data via Fully Nonparametric Bayesian Accelerated Failure Time Models*. 1–41.

Hirschfeld, R. M. A., Klerman, G. L., Gough, H. G., Barrett, J., Korchin, S. J., & Chodoff, P. (1977). A Measure of Interpersonal Dependency. *Journal of Personality Assessment*, *41*(6), 610–618. https://doi.org/10.1207/s15327752jpa4106_6

Holland, P. W. (1986). Statistics and causal inference. *Journal of the American Statistical Association*, *81*(396), 945–960.

Holzhauer, C. G., Epstein, E. E., Cohn, A. M., McCrady, B. S., Graff, F. S., & Cook, S. (2017).

Heterogeneity in Pathways to Abstinence among Women in Treatment for Alcohol Use Disorder. *Journal of Substance Abuse Treatment*, *75*, 1–9. https://doi.org/10.1016/j.jsat.2017.01.002.Heterogeneity

Holzhauer, C. G., Hildebrandt, T., Epstein, E., Mccrady, B., Hallgren, K. A., & Cook, S. (2020). Mechanisms of Change in Female-Specific and Gender-Neutral Cognitive Behavioral Therapy for Women With Alcohol Use Disorder. *Journal of Consulting and Clinical Psychology*, *88*(6), 541–553.

Huang, Y., Gilbert, P. B., & Janes, H. (2012). Assessing treatment-selection markers using a potential outcomes framework. *Biometrics*, *68*(3), 687–696. https://doi.org/10.1111/j.1541-0420.2011.01722.x

Imai, K., & Ratkovic, M. (2013). Estimating treatment effect heterogeneity in randomized program evaluation. *Annals of Applied Statistics*, *7*(1), 443–470. https://doi.org/10.1214/12-AOAS593

Imai, K., & Strauss, A. (2011). Estimation of heterogeneous treatment effects from randomized experiments, with application to the optimal planning of the get-out-the-vote campaign. *Political Analysis*, *19*(1), 1–19. https://doi.org/10.1093/pan/mpq035

Kadden, R., Carroll, K., Donovan, D., Cooney, N., Monti, P., Abrams, D., Litt, M., & Hester, R. (1995). Cognitive Behavioral Coping Skills Therapy Manual. *National Institute on Alcohol Abuse and Alcoholism Project MATCH Monograph Series*, *3*, 1–100.

Kapelner, A., Bleich, J., Levine, A., Cohen, Z. D., DeRubeis, R. J., & Berk, R. (2014). *Inference for the Effectiveness of Personalized Medicine with Software*. http://arxiv.org/abs/1404.7844

Kent, D. M., Steyerberg, E., & van Klaveren, D. (2018). Personalized evidence based medicine :

predictive approaches to heterogeneous treatment effects. *BMJ*, *364*, k4245. https://doi.org/10.1136/bmj.k4245

Kranzler, H. R., & McKay, J. R. (2012). Personalized Treatment of Alcohol Dependence. *Substance Use and Related Disorders*, *14*, 486–493. https://doi.org/10.1007/s11920-012-0296-5

Lamont, A., Lyons, M. D., Jaki, T., Stuart, E., Feaster, D. J., Tharmaratnam, K., Oberski, D., Ishwaran, H., Wilson, D. K., & Van Horn, M. L. (2016). Identification of predicted individual treatment effects in randomized clinical trials. *Statistical Methods in Medical Research*, 0962280215623981. https://doi.org/10.1177/0962280215623981

Lane, S. P., & Sher, K. J. (2015). Limits of current approaches to diagnosis severity based on criterion counts: An example with DSM-5 alcohol use disorder. *Clinical Psychological Science*, *3*, 819–835. https://doi.org/10.1177/2167702614553026

Litten, R. Z., Ryan, M. L., Falk, D. E., Reilly, M., Fertig, J. B., & Koob, G. F. (2015). Heterogeneity of Alcohol Use Disorder: Understanding Mechanisms to Advance Personalized Treatment. *Alcoholism: Clinical and Experimental Research*, *39*(4), 579–584. https://doi.org/10.1111/acer.12669

Longabaugh, R., Wirtz, P. W., Mattson, M. E., & Myers, J. K. (2001). Project MATCH hypotheses: Results and causal chain analyses. *National Institute on Alcohol Abuse and Alcoholism Project MATCH Monograph Series*, *8*, 1–341. http://pubs.niaaa.nih.gov/publications/ProjectMatch/match08.pdf%0D%0D%5Cn http://pubs.niaaa.nih.gov/publications/ProjectMatch/match08.pdf%5Cr%5Cr%5Cnhttp://pubs.niaaa.nih.gov/publications/ProjectMatch/match08.pdf#page=251

Magill, M., & Ray, L. A. (2009). Cognitive-Behavioral Treatment With Adult Alcohol and Illicit

Drug Users : A Meta-Analysis of Randomized Controlled Trials *. *Journal of Studies on Alcohol and Drugs*, *70*, 516–527.

Magura, S., Cleland, C. M., & Tonigan, J. S. (2013). Evaluating alcoholics anonymous's effect on drinking in project MATCH using cross-lagged regression panel analysis. *Journal of Studies on Alcohol and Drugs*, *74*(3), 378–385. https://doi.org/10.15288/jsad.2013.74.378

Maisto, S. A., Hallgren, K. A., Roos, C. R., & Witkiewitz, K. (2018). Course of remission from relapse to heavy drinking following outpatient treatment of alcohol use disorder. *Drug and Alcohol Dependence*, *187*, 319–326. https://doi.org/10.1016/j.drugalcdep.2018.03.011.Course

Maisto, S. A., Kirouac, M., & Witkiewitz, K. (2014). Alcohol use disorder clinical course research: informing clinicians' treatment planning now and in the future. *Journal of Studies on Alcohol and Drugs*, *75*(5), 799–807.

Maisto, S. A., Roos, C. R., O'Sickey, A. J., Kirouac, M., Connors, G. J., Tonigan, J. S., & Witkiewitz, K. (2015). The indirect effect of the therapeutic alliance and alcohol abstinence self-efficacy on alcohol use and alcohol-related problems in Project MATCH. *Alcoholism, Clinical and Experimental Research*, *39*(3), 504–513. https://doi.org/10.1111/acer.12649

Mann, K., & Hermann, D. (2010). Individualised treatment in alcohol-dependent patients. *European Archives of Psychiatry and Clinical Neuroscience*, *260*, S116–S120. https://doi.org/10.1007/s00406-010-0153-7

Mann, K., Roos, C. R., Hoffmann, S., Nakovics, H., Leménager, T., Heinz, A., & Witkiewitz, K. (2018). Precision medicine in alcohol dependence: A controlled trial testing pharmacotherapy response among reward and relief drinking phenotypes. *Neuropsychopharmacology*, *43*(4), 891–899. https://doi.org/10.1038/npp.2017.282

Mattson, M. E., Allen, J. P., Longabaugh, R., Nickless, C. J., Connors, G. J., & Kadden, R. M. (1994). A Chronological Review of Empirical Studies Matching to Treatment *. *Journal of Studies on Alcohol*, *Supp. No.*, 16–29.

McClelland, G. H., & Judd, C. M. (1993). Statistical Difficulties of Detecting Interactions and Moderator Effects. *Psychological Bulletin*, *114*(2), 376–390.

McKay, J. R., van Horn, D., Oslin, D. W., Ivey, M., Drapkin, M. L., Coviello, D. M., Yu, Q., & Lynch, K. G. (2011). Extended telephone-based continuing care for alcohol dependence: 24-month outcomes and subgroup analyses. *Addiction*, *106*(10), 1760–1769. https://doi.org/10.1111/j.1360-0443.2011.03483.x

McLellan, A. T., Luborsky, L., Woody, G. E., & O'Brien, C. P. (1980). An improved diagnostic evaluation instrument for substance abuse patients: The Addiction Severity Index. *Journal of Nervous and Mental Disease*, *168*(1), 26–33.

Miller, W. R. (1996). FORM 90: A Structured Assessment Interview for Drinking and Related Behaviors Test Manual. *NIAAA: Project MATCH Monograph Series*, *5*.

Miller, W., Zweben, A., DiClemente, C., & Rychtarik, R. (1995). Motivational Enhancement Therapy Manual. *National Institute on Alcohol Abuse and Alcoholism Project MATCH Monograph Series*, *2*, 1–122.

Poulson, R. S. (2011). Treatment heterogeneity and individual qualitative interaction. In *ProQuest Dissertations and Theses*. Kansas State University.

Powers, S., Qian, J., Jung, K., Schuler, A., Shah, N. H., Hastie, T., & Tibshirani, R. (2018). Some methods for heterogeneous treatment effect estimation in high dimensions. *Statistics in Medicine*, *37*, 1767–1787. https://doi.org/10.1002/sim.7623

Prochaska, J., & DiClemente, C. (1982). Transtheoretical therapy: Toward a more integrative model of change. *Psychotherapy*, *19*(3), 276–288. https://doi.org/10.1037/h0088437

Project Match Research Group. (1997). Matching alcoholism treatments to clients' heterogeneity: Project MATCH post-treatment drinking outcomes. *Journal of Studies on Alcohol*, *58*(1), 7–29.

Project MATCH Research Group. (1993). Rationale and Methods for a multisite clinical trial matching patients to alcoholism treatment. *Alcoholism: Clinical and Experimental Research*, *17*, 1130–1145.

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.

Ray, L. A., Meredith, L. R., Kiluk, B. D., Walthers, J., Carroll, K. M., & Magill, M. (2020). Combined Pharmacotherapy and Cognitive Behavioral Therapy for Adults With Alcohol or Substance Use Disorders A Systematic Review and Meta-analysis. *JAMA Network Open*, *3*(6), e208279. https://doi.org/10.1001/jamanetworkopen.2020.8279

Reitan, R. M. (1958). Validity of the Trail Making Test as an indicator of organic brain damage. *Perceptual and Motor Skills*, *8*(3), 271–276.

Rekkas, A., Paulus, J. K., Raman, G., Wong, J. B., Steyerberg, E. W., Rijnbeek, P. R., Kent, D. M., & Klaveren, D. van. (2019). Predictive approaches to heterogeneous treatment effects: a systematic review. *MedRxiv*, 19010827. https://doi.org/10.1101/19010827

Roos, C. R., Mann, K. F., & Witkiewitz, K. (2017). Reward and relief dimensions of temptation to drink: Construct validity and role in predicting differential benefit from acamprosate and naltrexone. *Addiction Biology*, *22*, 1528–1539. https://doi.org/10.1111/adb.12427

Rosenbaum, P. R. (1984). Conditional permutation tests and the propensity score in observational studies. *Journal of the American Statistical Association*, *79*(387), 565–574. https://doi.org/10.1080/01621459.1984.10478082

Ruberg, S. J., Chen, L., & Wang, Y. (2010). The mean does not mean as much anymore: finding sub-groups for tailored therapeutics. *Clinical Trials (London, England)*, *7*(5), 574–583. https://doi.org/10.1177/1740774510369350

Rubin, D. B. (2005a). Causal Inference Using Potential Outcomes: Design, Modeling, Decisions. *Journal of the American Statistical Association*, *100*, 322–331.

Rubin, D. B. (2005b). Causal Inference Using Potential Outcomes Causal Inference Using Potential Outcomes : Design , Modeling , Decisions. *Journal of the American Statistical Association*, *100*(469), 322–331. https://doi.org/10.1198/016214504000001880

Sarason, I. G., Levine, H. M., Basham, R. B., & Sarason, B. R. (1983). Assessing social support: The Social Support Questionnaire. *Journal of Personality and Social Psychology*, *44*(1), 127–139. https://doi.org/10.1037/0022-3514.44.1.127

Schafer, J. L. (1999). Multiple imputation: A primer. *Statistical Methods in Medical Research*, *8*(1), 3–15. https://doi.org/10.1191/096228099671525676

Seibold, H., Zeileis, A., & Hothorn, T. (2018). *Individual treatment effect prediction for amyotrophic lateral sclerosis patients*. https://doi.org/10.1177/0962280217693034

Shen, C., Jeong, J., Li, X., Chen, P.-S., & Buxton, A. (2013). Treatment benefit and treatment harm rate to characterize heterogenetiy in treatment effect. *Biometrics*, *69*(3), 724–731. https://doi.org/10.1111/biom.12038.Treatment

Shipley, W. C. (1940). A self-administering scale for measuring intellectional impairment and deterioration. *Journal of Psychology*, *9*, 371–377.

Sliedrecht, W., de Waart, R., Witkiewitz, K., & Roozen, H. G. (2019). Alcohol use disorder relapse factors: A systematic review. *Psychiatry Research*, *278*, 97–115. https://doi.org/10.1016/j.psychres.2019.05.038

Smith, A. (1973). *Symbol Digit Modalities Test*. Western Psychological Services.

Smith, M. D., Saunders, R. S., Stuckhardt, L., & McGinnis, J. M. (2013). *Best care at lower cost: The path to continuously learning health care in America* (Institute of Medicine (Ed.)). National Academies Press.

Strobl, C., Boulesteix, A.-L., Kneib, T., Augustin, T., & Zeileis, A. (2008). Conditional variable importance for random forests. *BMC Bioinformatics*, *9*, 307–318. https://doi.org/10.1186/1471-2105-9-307

Tomko, R. L., Bountress, K. E., & Gray, K. M. (2016). Personalizing substance use treatment based on pre-treatment impulsivity and sensation seeking: A review. *Drug and Alcohol Dependence*, *167*, 1–7. https://doi.org/10.1016/j.drugalcdep.2016.07.022

Tonigan, J. S., Connors, G. J., & Miller, W. R. (1996). Alcoholics Anonymous Involvement (AAI) Scale: Reliability and Norms. *Psychology of Addictive Behaviors*, *10*(2), 75–80.

Van Der Stel, J. (2015). Precision in Addiction Care: Does it Make a Difference? *Yale Journal of Biology and Medicine*, *88*, 415–422.

Volkow, N. D. (2020). Personalizing the Treatment of Substance Use Disorders. *American Journal of Psychiatry*, *177*(2), 113–116. https://doi.org/10.1176/appi.ajp.2019.19121284

Waddington, L. (2002). The therapy relationship in cognitive therapy: A review. *Behavioural and Cognitive Psychotherapy*, *30*(2), 179–191. https://doi.org/10.1017/S1352465802002059

Webb, C. A., Cohen, Z. D., Beard, C., Forgeard, M., Peckham, A. D., & Björgvinsson, T. (2020). Personalized Prognostic Prediction of Treatment Outcome for Depressed Patients in a Naturalistic Psychiatric Hospital Setting : A Comparison of Machine Learning

Approaches. *Journal of Consulting and Clinical Psychology*, *88*(1), 25–38.

Witkiewitz, K., Hartzler, B., & Donovan, D. (2010). Matching motivation enhancement treatment to client motivation: Re-examining the Project MATCH motivation matching hypothesis. *Addiction*, *105*(8), 1403–1413. https://doi.org/10.1111/j.1360-0443.2010.02954.x

Witkiewitz, K., Litten, R. Z., & Leggio, L. (2019). Advances in the science and treatment of alcohol use disorder. *Science Advances*, *5*, eaax4043.

Witkiewitz, K., Maisto, S. A., & Donovan, D. M. (2010). A comparison of methods for estimating change in drinking following alcohol treatment. *Alcoholism: Clinical and Experimental Research*, *34*(12), 2116–2125. https://doi.org/10.1111/j.1530-0277.2010.01308.x.A

Witkiewitz, K., & Marlatt, G. A. (2004). Relapse prevention for alcohol and drug problems: That was Zen, this is Tao. *American Psychologist*, *59*(4), 224–235. https://doi.org/10.1037/0003-066X.59.4.224

Witkiewitz, K., Roos, C. R., Mann, K., & Kranzler, H. R. (2019). Advancing precision medicine for alcohol use disorder: Replication and extension of reward drinking as a predictor of naltrexone response. *Alcoholism: Clinical and Experimental Research*, *43*(11), 2395–2405. https://doi.org/10.1111/acer.14183

Witkiewitz, K., Wilson, A. D., Pearson, M. R., Montes, K. S., Kirouac, M., Roos, C. R., Hallgren, K. A., & Maisto, S. A. (2019). Profiles of recovery from alcohol use disorder at three years following treatment: can the definition of recovery be extended to include high functioning heavy drinkers? *Addiction*, *114*(1), 69–80. https://doi.org/10.1111/add.14403

Zhang, Z., Wang, C., Nie, L., & Soon, G. (2013). Assessing the heterogeneity of treatment effects via potential outcomes of individual patients. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, *62*(5), 687–704.

Table 1

Descriptive Statistics of Baseline MATCH Covariates and Abstinence

	Outpatient			Aftercare			Chi-Square/ t-test
	Mean/ proportion	SD	Range	Mean/ proportion	SD	Range	X^2/t(p-value)
Baseline Covariates							
Female	0.27	0.44	0.0-1.0	0.21	0.41	0.0-1.0	2.26 (0.02)
Age	38.62	10.66	18.0-73.0	42.20	11.44	20.0-76.0	72.95 (0.04)
Cognitive Functioning:							
Shipley-Hartford (CFS)	-0.43	2.14	-5.9-7.2	0.51	2.67	-5.5-17.2	1123.9 (0.77)
State-Trait Anger Scale (TAS)	29.60	7.41	15.0-57.0	30.87	7.43	15.0-56.0	59.11 (0.04)
Religious Background and Beliefs (RBB)	34.97	10.82	13.0-67.0	38.77	11.52	16.0-71.0	74.32 (0.03)
Psychosocial Functioning Inventory (PFI)	0.51	0.17	0.03-0.98	0.45	0.17	0.02-1.00	134.11 (0.01)
Alcohol Abstinence Self-Efficacy Scale-Confidence (AASE-C)	2.96	0.80	1.0-5.0	3.24	1.03	1.0-5.0	163.32 (0.00)

Alcohol Abstinence Self-Efficacy Scale-Temptation (AASE-T)	0.05	1.38	-4.0-3.7	-0.46	1.67	-4.0-3.8	270.79 (0.00)
Interpersonal Dependency Instrument (IDI)	41.13	7.16	19.0-56.0	41.01	7.43	17.0-56.0	42.50 (0.28)
Consequences of Drinking Scale	3.96	0.19	3.0-4.0	3.98	0.15	2.0-4.0	187.84 (0.00)
Number of Drinks per Drinking Day	3.53	0.98	1.6-7.6	4.30	1.27	1.7-7.8	1136.50 (0.38)
Percent Days Abstinent	0.57	0.38	0.0-1.5	0.45	0.40	0.0-1.5	130.22 (0.02)
Addiction Severity Index (ASI) --							
Psychiatric Status	0.19	0.19	0.0-0.8	0.23	0.21	0.0-0.8	214.38 (0.28)
Alcoholics Anonymous Involvement (AAI)	0.93	0.42	0.6-2.0	1.32	0.41	0.0-2.0	243.05 (0.00)
Social Supports-Family	3.52	1.54	0.0-7.0	3.42	1.53	0.0-7.0	3.59 (0.83)
Social Supports-Friends	3.16	1.48	0.0-7.0	7.00	1.62	0.0-6.0	15.49 (0.03)
Readiness to Change (URICA)	10.53	1.70	3.0-14.0	11.07	1.53	1.6-14.0	181.32 (0.00)
Alcoholism Typology	0.85	0.67	0.0-2.0	0.97	0.69	0.0-2.0	9.99 (0.01)
Seeking of Noetic Goals (SONG)	78.27	17.47	22.0-131.0	81.76	18.52	31.0-136.0	97.66 (0.42)
Outcome (15 months)							
Abstinence	0.37	0.48	0-1	0.56	0.50	0-1	-2.43 (0.00)

Table 2

Logistic Regression Predicting Abstinence from Baseline Characteristics

	Aftercare (n=527)						Outpatient (n=617)					
	CBT (n=266)			MET (n=216)			CBT (n=301)			MET (n=316)		
	b		SE	b		SE	b		SE	b		SE
Intercept	3.77		3.68	0.68		1.99	-7.56	*	3.45	1.83		3.47
Baseline Covariates												
Female	0.33		0.36	0.70	*	0.38	-0.71	*	0.33	0.27		0.30
Age	0.03	†	0.01	0.02	†	0.01	-0.00		0.01	-0.00		0.01
Cognitive Functioning: Shipley-Hartford (CFS)	-0.00		0.06	-0.08		0.07	-0.01		0.06	-0.07		0.06
State-Trait Anger Scale (TAS)	-0.03		0.02	0.02		0.02	0.00		0.02	0.01		0.02
Religious Background and Beliefs (RBB)	0.02		0.01	0.00		0.01	0.02	†	0.01	0.00		0.01
Psychosocial Functioning Inventory (PFI)	-2.41	*	1.09	-1.58	†	0.90	1.90	†	1.03	-0.83		0.91
Alcohol Abstinence Self-Efficacy Scale-Confidence (AASE-C)	-0.21		0.24	-0.13		0.23	0.51		0.35	-0.53	†	0.30
Alcohol Abstinence Self-Efficacy Scale-Temptation (AASE-T)	-0.36	*	0.16	-0.08		0.15	0.15		0.22	-0.24		0.19
Interpersonal Dependency Instrument (IDI)	0.01		0.02	-0.01		0.02	0.02		0.02	-0.04	*	0.02

Consequences of Drinking Scale	-1.05		0.77	-[a]		-[a]	0.33		0.64	-0.21		0.75
Number of Drinks per Drinking Day	0.07		0.13	0.13		0.13	0.01		0.15	0.23	†	0.13
Percent Days Abstinent	0.92	*	0.36	0.55		0.39	0.41		0.37	-0.10		0.35
Addiction Severity Index (ASI) Psychiatric Status	-0.45	†	0.71	-2.11	**	0.74	0.46		0.73	-0.78		0.74
Alcoholics Anonymous Involvement (AAI)	-0.69	†	0.37	0.24		0.36	-0.06		0.34	0.36		0.32
Social Supports-Family	0.09		0.10	-0.16		0.09	-0.17	†	0.09	0.02		0.08
Social Supports-Friends	-0.02		0.09	-0.05		0.09	-0.11		0.09	-0.08		0.09
Readiness to Change (URICA)	0.12		0.11	-0.03		0.08	0.23	**	0.08	0.14		0.09
Type A versus B Alcoholism	0.04		0.34	-0.00		0.31	0.74	*	0.33	0.01		0.30
Seeking of Noetic Goals (SONG)	-0.00		0.01	-0.00		0.01	-0.01		0.01	-0.01		0.01

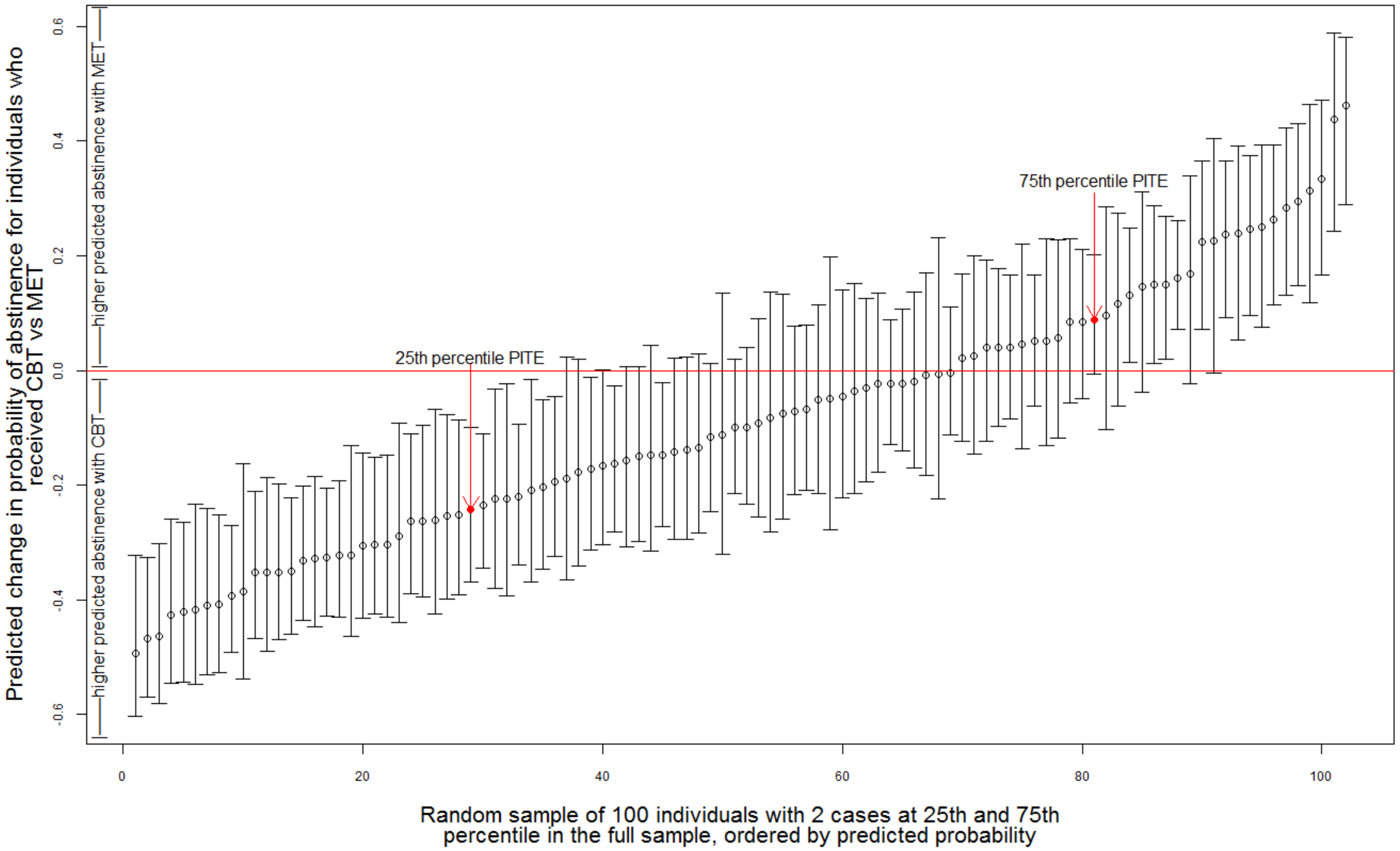
Note: P-values refer to the significance of the regression weights for the predictive models.

† p < .10; * p < .05; ** p < 0.01

[a] Variable for consequences of drinking was eliminated from the model for participants in the aftercare sample receiving MET due to a lack of variability in the variable among this group.

Figure 1

Predicted Effects of CBT versus MET for a Random Sample of 100 Individuals from the Outpatient Sample of Project Match



Note: The dots represent the predicted difference in the probability of abstinence for each person under CBT and MET. The horizontal line is where the two intervention effects are equal. The lower interval for each dot indicates that the individual is expected to be at or above the lowest point 80% of the time and the upper interval indicates that the individual is expected to at or below the highest point 80% of the time. If the interval does not include zero it would suggest good support for CBT or MET being better for that individual.

Appendix A: Data Transparency

The data reported in this manuscript have been previously published and were collected as part of a larger data collection. Findings from the data collection have been reported in separate manuscripts. Other studies using MATCH data have looked at a large array of average treatment effects, including the specific hypothesized a priori interactions using the study's primary and secondary matching hypotheses and many of the other psychosocial processes that underlay alcohol misuse and treatment. The current study differs from and extends on previous research by using a personalized medicine approach to simultaneously test whether there is evidence for differential treatment effects when a large number of individual level variables are simultaneously used to test for differential effects of the treatment. We know of no previous study which has used a similar analytic approach to test for differences in the treatments used in the MATCH study.

Appendix B: Algorithm for Deriving PITE Predictions

```
#Import imputed CBT treatment group
cbt <- read.csv("M:/Project MATCH/mi_cbtimpFINAL.csv")

#select first imputed dataset
cbt <- cbt[which(cbt$.imp==1),]
colnames(cbt)

#Import imputed MET treatment group
met <- read.csv("M:/Project MATCH/mi_metimpFINAL.csv")

#select first imputed dataset
met <- met[which(met$.imp==1),]
colnames(met)

match <- rbind(cbt, met)                        #Combine datasets
dim(match)                                      #check # of rows and columns

match$txassgn<-match$txassgn-1            #Recode treatment variable as 0/1
table(match$txassgn)                            ###met=1, cbt=0

#remove extraneous outcomes/non-baseline vars that had been used in imputation
model
match<-match[,c(-36,-37,-40:-43,-108)]

#remove extraneous vars created during imputation
match[,c(1:3)]<-NULL

###All changes saved in "M:/Project Match/Data/finalpitematch.csv"###
match<-read.csv("M:/Project Match/Data/finalpitematch.csv")

########################Main Effect of Treatment###########################
library(lme4)

#Subset sample into outpatient and aftercare
match.out <- match[which(match$arm==1),]
match.aft <- match[which(match$arm==2),]

##For Table 2
#Outpatient
summary(m1<-glm(abstinent~Female + age + cogimp_CFS + angers_TAS + rbbtots +
socfuncs_PFI + selfef1C_AASE + selfef2T_AASE + idiscors + otcmi_Tot +
ydrkr0_Tot + ypror0_Tot +  asipsya_Tot + aai_int_Tot + ss2Famtot + ss1frietot
+ urica + typol + song_a, data=match.out.cbt, family="binomial"))
summary(m2<-glm(abstinent~Female + age + cogimp_CFS + angers_TAS + rbbtots +
socfuncs_PFI + selfef1C_AASE + selfef2T_AASE + idiscors + otcmi_Tot +
ydrkr0_Tot + ypror0_Tot +  asipsya_Tot + aai_int_Tot + ss2Famtot + ss1frietot
+ urica + typol + song_a, data=match.out.met, family="binomial"))
#Aftercare
summary(m3<-glm(abstinent~Female + age + cogimp_CFS + angers_TAS + rbbtots +
socfuncs_PFI + selfef1C_AASE + selfef2T_AASE + idiscors + otcmi_Tot +
ydrkr0_Tot + ypror0_Tot +  asipsya_Tot + aai_int_Tot + ss2Famtot + ss1frietot
+ urica + typol + song_a, data=match.aft.cbt, family="binomial"))
summary(m4<-glm(abstinent~Female + age + cogimp_CFS + angers_TAS + rbbtots +
socfuncs_PFI + selfef1C_AASE + selfef2T_AASE + idiscors + otcmi_Tot +
ydrkr0_Tot + ypror0_Tot +  asipsya_Tot + aai_int_Tot + ss2Famtot + ss1frietot
+ urica + typol + song_a, data=match.aft.met, family="binomial"))

####################PITE & PERMUTATION TEST, BY ARM#########################
colnames(match.out)
out.cov<-match.out[,c("Female","age","cogimp_CFS","angers_TAS","rbbtots",
      #create matrix with baseline covariates among outpatient sample
      "socfuncs_PFI","selfef1C_AASE","selfef2T_AASE","idiscors",
      "otcmi_Tot","ydrkr0_Tot","ypror0_Tot","asipsya_Tot","aai_int_Tot",
      "ss2Famtot","ss1frietot","urica","song_a","typol")]

out.pite<-PiteBinary(match.out$txassgn,match.out$abstinent,out.cov)
            #PITE function
str(out.pite)
            #elements of results object
set.seed(9359503)
out.perm<-Pite.prmtYasin(out.pite,1000,responseType = 'binary')
            #Permutation test function using PITE results, 1000 permutations
for binary outcome
str(out.perm)
            #elements of permutation test results object
##p=0.024

match.aft <- match[which(match$arm==2),]
      #subset aftercare sample
colnames(match.aft)
aft.cov<-match.aft[,c("Female","age","cogimp_CFS","angers_TAS","rbbtots",
      #create matrix with baseline covariates among aftercare sample
      "socfuncs_PFI","selfef1C_AASE","selfef2T_AASE","idiscors",
      "otcmi_Tot","ydrkr0_Tot","ypror0_Tot","asipsya_Tot","aai_int_Tot",
      "ss2Famtot","ss1frietot","urica","song_a","typol")]

aft.pite<-PiteBinary(match.aft$txassgn,mat.aft$abstinent,aft.cov)
      #PITE function
str(aft.pite)
      #elements of results object
set.seed(9359223)
aft.perm<-Pite.prmtYasin(aft.pite,1000,responseType = 'binary')
      #Permutation test function using PITE results, 1000 permutations for
binary outcome
str(aft.perm)
      #elements of permutation test results object
##p=0.277



##***********************PITE PREDICT FUNCTION***************************##
##PITE function for use with binary outcome##
PiteBinary <- function(trt, y, cov, inter = F, extract='ATE', covToInc='all'){
  n <- nrow(cov)
  NumCov <- dim(cov)[2]
  if (length(table(trt)) != 2)
    stop("The treatment level is not two.")

#Code is designed to test treatment vs. control group, if there are too many
levels in the designated treatment variable, function will stop and produce
this error
```

```
  if( inter == TRUE & n/2 <= (factorial(ncol(cov)) / (factorial(ncol(cov)-
2)*factorial(2)) + ncol(cov) +1) )
    stop("degrees of freedom of residual is negative, sample size (n) needs to
be larger ")
  if( inter == FALSE & n/2 <= (ncol(cov) +1) )
    stop("degrees of freedom of residual is negative, sample size (n) needs to
be larger
         than twice of the number of total covariates plus 1 (NumCov +
NumNuiCov +1)")

#Default is set to not consider interactive effects, but if set to inter=TRUE,
they will be included in the calculation of the PITEs,
#if sample size is not large enough for adequate degrees of freedom, this
error will be returned

  if(extract !='model' & extract !='ATE'){
    stop("extraction method is not ATE or model")
  }

#Default is set to extract the average treatment effect of the intervention,
so that PITEs are calculated as individual treatment effects
#above and beyond the average effect of the intervention

  if(covToInc !='all' & length(covToInc)> ncol(cov)){
    stop("Number of covariates to include is greater than given number of
covariates")     }


#============================================================================#

  # Set up data into required format
  data <- data.frame(y, trt)
  dataxy <- data.frame(y, cov)
  # create the interaction between pair predictors x, add in to the dataset
  # dataint now have y, treatment condition, x, and the pairwise interaction
between x (if applicable)
  options(na.action = 'na.pass')
  dataint2 <- model.matrix(y ~ .^2, data = dataxy)    #creates matrix of each
case's coefficients for baseline covariates and pairwise interactions
  dataint <- data.frame(y, trt, x = dataint2[,-1])    #creates data frame of
y, treatment condition, and covariate + pairwise interactions (minus
intercept)


  # split covariates based on treatment assignment
  xt <- dataint[dataint$trt ==1, 3:ncol(dataint)]
  xc <- dataint[dataint$trt ==0, 3:ncol(dataint)]

  # split responses based on treatment assignment and order
  yt <- y[dataint$trt == 1]
  yc <- y[dataint$trt == 0]
  y.ord <- c(yt,yc)

  # separate the dataset (dataint) to treatment group and control group
  TRTmxi <- data.frame(xt)
  CNTLmxi <- data.frame(xc)

  # assigns treatment labels for ordered data
  trt1<- c(rep(1,length(yt)), rep(0, length(yc)))

    if(extract == 'model'){
            if(covToInc =='all'){ # set up which covariates to use for this 
model
                daitaint1 <- dataint
              }
            else{
                cov1 <- cov[,covToInc]
                dataint2.1 <- model.matrix(y ~ .^2, data =data.frame(y,cov1))
                dataint1 <- data.frame(y, trt, x = dataint2.1[,-1])
            }

        # extract coefficients from the control model
         mod <- glm(dataint1[,1] ~., family =binomial(link = 'logit'), data = 
dataint1[,-1])
        lambda <- coefficients(mod)[2]
    }

    else if(extract=='ATE'){
        # simple model
        mod <- glm(y.ord ~trt1, family =binomial(link = 'logit'))
        lambda <- coefficients(mod)[2]
    }

  # if the interaction of predictors are not taken into consideration
  if (inter == FALSE) {

    # fit model for trt group
    mod.t <- glm(yt ~ .,family =binomial(link = 'logit'), data = TRTmxi[,1:
(ncol(cov))])
    # fit model for control group
    mod.c <- glm(yc ~ .,family =binomial(link = 'logit'), data = CNTLmxi[,1:
(ncol(cov))])
    et <- coefficients(mod.t)
    ec <- coefficients(mod.c)
    cov <- as.matrix(cov)

    # predictions of response on logit scale
    pred_t <- et[1] + (cov%*%et[2:(length(et))])
    pred_c <- ec[1] + (cov%*%ec[2:(length(ec))])

    # predictions of responses on probability scale
    ppPred_t <- sapply(pred_t, logistic)
    ppPred_c <- sapply(pred_c, logistic)

    # FOR PERMUTATION TEST first on logit scale followed by probability scale;
    predt.perm <- pred_t - lambda                     #subtract out average 
treatment effect (ATE) from individual effects
```

```
    ppPredt.perm <- sapply(predt.perm, logistic)      #convert to probability
scale
  }

  # if the pairwise interaction of predictors are taken into consideration
  if (inter == TRUE) {

    mod.t <- glm(yt ~ .,family =binomial(link = 'logit'), data = TRTmxi) # fit
model for trt group
    mod.c <- glm(yc ~ .,family =binomial(link = 'logit'), data = CNTLmxi) #
fit model for control group
    et <- coefficients(mod.t)
    ec <- coefficients(mod.c)
    cov <- as.matrix(cov)
    # predictions of response on trt on logit scale
    pred_t <- et[1] + (dataint[,3:ncol(dataint)]%*%et[2:(length(et))])
    pred_c <- ec[1] + (dataint[,3:ncol(dataint)]%*%ec[2:(length(ec))])

    ppPred_t <- sapply(pred_t, logistic)   #predictions on probability scale
    ppPred_c <- sapply(pred_c, logistic)   #predictions on probability scale

    # FOR PERMUTATION TEST first on logit scale followed by probability scale;
    predt.perm <- pred_t - lambda
    ppPredt.perm <- sapply(predt.perm, logistic)

  }

  # calculate the estimate: individual PITE estimate and the standard
deviation
  # of PITE estimates among individuals
  pite.ind <- ppPred_t - ppPred_c         #predicted response under treatment
condition - predicted response under control
  pite.sd <- sd(pite.ind)

  pite.ind.perm <-  ppPredt.perm - ppPred_c
  pite.sd.perm <- sd(pite.ind.perm)

  return(list("trt.coef" = et,
              "cntl.coef" = ec,
              "pite.trt.cond" = ppPred_t,
              "pite.cntl.cond" = ppPred_c,
              "pite.trt.PT" =ppPredt.perm,
              "pite.ind" = pite.ind,
              "pite.sd" = pite.sd,
              "pite.ind.perm" =pite.ind.perm,
              "pite.sd.perm" = pite.sd.perm,
              "TOTCov" = ncol(cov),
              "gen.data" = data,
              "gen.data.int" = dataint,
              "NumCov" = NumCov,
              "gen.xMain" = cov,
              "gen.xMainInt" = dataint[,3:ncol(dataint)],
              "interaction" = inter,
              "TOTCov" = ncol(cov)
  ))
}


#========================PERMUTATION FUNCTION============================#

Pite.prmtYasin <- function(pite.rslt, nperm, inter = F, responseType='normal',
extract = 'ATE', covToInc='all'){
  if(covToInc !='all' & length(covToInc)> ncol(pite.rslt$gen.xMain)){
    stop("Number of covariates to include is greater than given number of
covariates")     }
  if(responseType!='binary' & responseType !='normal'){
    stop("Response type is not normal or binary")
  }
  if(extract !='model' & extract !='ATE'){
    stop("extraction method is not ATE or model")
  }



  # use the output from 'pite.rslt' to know where the PITE SD lies for
original treatment labels
  # use output which has subtracted treatment effect
  pite.ind <- pite.rslt$pite.ind.perm
  pite.sd <- pite.rslt$pite.sd.perm
  trt <- pite.rslt$gen.data$trt # original treatment labels
  y <- pite.rslt$gen.data$y  # observed responses

  # setting up covariates if interactions F, only covariates, if T, include
interactions
  if (inter == F) {
    x <- pite.rslt$gen.xMain
  }

  if (inter == T) {
    x <- pite.rslt$gen.xMainInt
  }

  # set up a matrix with all permutations of treatment assignment.
  # each column corresponds to one set of permutated labels.
  trtperms <- replicate(n=nperm, sample(pite.rslt$gen.data$trt, replace = F))

 # dependent on response type, we apply the correct function: Normal or Binary
  if( responseType == 'normal'){
    prmt <- sapply(apply(trtperms,2,PiteNormalPRMT,trt=trt,cov=x,y=y,
                         extract=extract, inter=inter, covToInc=covToInc),
'[[', 'pite.ind')
  }
#prmt=pites for each permutation
  else if(responseType == 'binary'){
    prmt <- sapply(apply(trtperms,2,PiteBinaryPRMT,trt=trt,cov=x,y=y,
                         extract=extract, inter=inter, covToInc=covToInc),
'[[', 'pite.ind')
  }

#calculating SD for each permutation
```

```
  prmt.sd <- apply(prmt, 2, sd)

#p value of permutation test calculated based on proportion of permuted SDs
that are greater than observed SD
#divided by total number of permutations
  alpha <-  sum(pite.sd <=prmt.sd)/nperm

# histogram of permuted SDs, line for SD of PITE in observed data (figure 2)
  hist(prmt.sd, xlim = c(min(pite.sd, range(prmt.sd)[1]) -.05, max(pite.sd,
range(prmt.sd)[2])+.05),
       main = (deparse(substitute(pite.rslt))), sub = paste0('p-value = ',
alpha))
  abline(v=pite.sd, col=2)

  return(list(
    "pite.ind" = pite.ind,
    "permute" = prmt,
    "pite.sd" = pite.sd,
    "prmt.sd" = prmt.sd,
    "alpha" = alpha
  ))
}


#=========== EXTRA FUNCTIONS REQUIRED FOR PERMUATION TEST ===============#
#This function randomly assigns treatment condition for as many times as
specified by nperm
#and then runs the original function to produce PITEs with these random
assignments
#Those results are collated and returned by Pite.prmtYasin

PiteBinaryPRMT <- function(trt1, trt, y, cov, inter = F, extract = 'ATE',
covToInc = 'all'){
  if(extract !='model' & extract !='ATE'){
    stop("extract method is not ATE or model")
  }
       n <- nrow(cov)
  NumCov <- dim(cov)[2]
  if (length(table(trt)) != 2)
    stop("The treatment level is not two.")
  if( inter == TRUE & n/2 <= (factorial(ncol(cov)) / (factorial(ncol(cov)-
2)*factorial(2)) + ncol(cov) +1) )
    stop("degrees of freedom of residual is negative, sample size (n) needs to
be larger ")
  if( inter == FALSE & n/2 <= (ncol(cov) +1) )
    stop("degrees of freedom of residual is negative, sample size (n) needs to
be larger than twice
         of the number of total covariates plus 1 (NumCov + NumNuiCov +1)")
  if(covToInc !='all' & length(covToInc)> ncol(cov)){
    stop("Number of covariates to include is greater than given number of
covariates")     }
#=============================================================================#

  data <- data.frame(y, trt)
    dataxy <- data.frame(y, cov)
  # create the interaction between pair predictors x, add in to the dataset
  # dataint now have y, treatment condition, x, and the pairwise interaction
between x

  options(na.action = 'na.pass')
  dataint2 <- model.matrix(y ~ .^2, data = dataxy)
  dataint <- data.frame(y, trt, x = dataint2[,-1])

  # SPLIT DATASETS ACCORDING TO PERMUTED TREATMENT LABELS
##This is where the permutation test code differs from the PITE function above
##Here, the dataset is being split up according to randomly /permuted/
treatment condition (trt1)
##Not according to observed treatment assignment (above as "trt")

  xt1 <- dataint[trt1 ==1, 3:ncol(dataint)] # covariates of treatment group
  xc1 <- dataint[trt1 ==0, 3:ncol(dataint)] # covariates of control group

  yt1 <- y[trt1 == 1]       #responses on treatment
  yc1 <- y[trt1 == 0]       #responses on control
  y.ord1 <- c(yt1,yc1)

  TRTmxi1 <- data.frame(xt1)
  CNTLmxi1 <- data.frame(xc1)

  if (extract == 'model'){
    if(covToInc =='all'){
      dataint1 <- dataint
    }
    else{
      cov1 <- cov[,covToInc]
      dataint2.1 <- model.matrix(y ~ .^2, data =data.frame(y,cov1))
      dataint1 <- data.frame(y, trt, x = dataint2.1[,-1])
    }

  # extract coefficients from the control model
  mod <- glm(dataint[,1] ~., family =binomial(link = 'logit'), data =
dataint[,-1])
  lambda <- coefficients(mod)[2]
  }
  else if(extract == 'ATE'){
  mod <- glm(dataint[,1] ~dataint[,2], family =binomial(link = 'logit'))
  lambda <- coefficients(mod)[2]
  }

  # finding those patients who are trt=1 in the group of people who are trt1=0
and trt1=1
  offset.t <- trt1[trt==1]
  offset.c <- trt1[trt==0]

  # if the interaction of predictors are not taken into considertion
  if (inter == FALSE) {
    # set up model on responses from permuted treatment labels including
offset
    mod.t <- glm(yt1 ~ .,family =binomial(link = 'logit'),
                 data = TRTmxi1[,1:(ncol(cov))], offset = lambda*offset.t)
    mod.c <- glm(yc1 ~ .,family =binomial(link = 'logit'),
                 data = CNTLmxi1[,1:(ncol(cov))], offset = lambda *offset.c)
    et <- coefficients(mod.t)
    ec <- coefficients(mod.c)
    cov <- as.matrix(cov)
    # predictions of responses on logit scale

    xbeta_t <- et[1] + (cov%*%et[2:(length(et))])
    xbeta_c <- ec[1] + (cov%*%ec[2:(length(ec))])

    # predictions of response on probability scale
    ppPred_t <- sapply(xbeta_t, logistic)
    ppPred_c <- sapply(xbeta_c, logistic)

  }

  # if the pairwise interaction of predictors are taken into consideration
  if (inter == TRUE) {

    # set up model on responses from permuted treatment labels including
offset
    mod.t <- glm(yt1 ~ .,family =binomial(link = 'logit'), data = TRTmxi1,
                 offset = lambda*offset.t)
    mod.c <- glm(yc1 ~ .,family =binomial(link = 'logit'), data = CNTLmxi1,
                 offset = lambda*offset.c)
    et <- coefficients(mod.t)
    ec <- coefficients(mod.c)
    dataint <- as.matrix(dataint)
    # predictions of responses on logit scale
    xbeta_t <- et[1] + (dataint[,3:ncol(dataint)]%*%et[2:(length(et))])
    xbeta_c <- ec[1] + (dataint[,3:ncol(dataint)]%*%ec[2:(length(ec))])

    #predictions on probability scale
    ppPred_t <- sapply(xbeta_t, logistic)
    ppPred_c <- sapply(xbeta_c, logistic)

  }

  # calculate PITE estimate and the standard deviation of pite estimates
  pite.ind <- ppPred_t - ppPred_c
  pite.sd <- sd(pite.ind)


  return(list("pite.trt.cond" = ppPred_t,
              "pite.cntl.cond" = ppPred_c,
              "pite.ind" = pite.ind,
              "pite.sd" = pite.sd,
              "TOTCov" = ncol(cov),
              "gen.data" = data,
              "gen.data.int" = dataint,
              "NumCov" = NumCov,
              "gen.xMain" = cov,
              "gen.xMainInt" = dataint[,3:ncol(dataint)],
              "interaction" = inter,
              "TOTCov" = ncol(cov)
  ))
}

#Function for converting estimates on the logit scale to the probability scale
logistic <- function(x){
  return(exp(x)/(1+exp(x)))
}
```