Received: 29 September 2023 | Revised: 11 July 2024 | Accepted: 16 July 2024

DOI: 10.1002/sim.10187

RESEARCH ARTICLE

Statistics in Medicine WILEY

Validation of predicted individual treatment effects in out of sample respondents

Alena Kuhlemeier[1] | Thomas Jaki[2,3] | Katie Witkiewitz[1] | Elizabeth A. Stuart[4] | M. Lee Van Horn[5]

[1]Center on Alcohol, Substance Use, and Addictions, University of New Mexico, Albuquerque, New Mexico, USA

[2]Chair for Computational Statistics, University of Regensburg, Regensburg, Germany

[3]MRC Biostatistics Unit, University of Cambridge, Cambridge, UK

[4]Department of Biostatistics, Johns Hopkins Bloomberg School of Public Health, Baltimore, Maryland, USA

[5]Department of Individual, Family, & Community Education, College of Education and Human Sciences, University of New Mexico, Albuquerque, New Mexico, USA

Correspondence
M. Lee Van Horn, Department of Individual, Family, & Community Education, College of Education and Human Sciences, University of New Mexico, Albuquerque, NM 87131, USA.
Email: mlvh@unm.edu

Funding information
National Institute on Alcohol Abuse and Alcoholism, Grant/Award Numbers: R01AA030264, T32AA018108, R01AA022328

Personalized medicine promises the ability to improve patient outcomes by tailoring treatment recommendations to the likelihood that any given patient will respond well to a given treatment. It is important that predictions of treatment response be validated and replicated in independent data to support their use in clinical practice. In this paper, we propose and test an approach for validating predictions of individual treatment effects with continuous outcomes across samples that uses matching in a test (validation) sample to match individuals in the treatment and control arms based on their predicted treatment response and their predicted response under control. To examine the proposed validation approach, we conducted simulations where test data is generated from either an identical, similar, or unrelated process to the training data. We also examined the impact of nuisance variables. To demonstrate the use of this validation procedure in the context of predicting individual treatment effects in the treatment of alcohol use disorder, we apply our validation procedure using data from a clinical trial of combined behavioral and pharmacotherapy treatments. We find that the validation algorithm accurately confirms validation and lack of validation, and also provides insights into cases where test data were generated under similar, but not identical conditions. We also show that the presence of nuisance variables detrimentally impacts algorithm performance, which can be partially reduced though the use of variable selection methods. An advantage of the approach is that it can be widely applied to different predictive methods.

KEYWORDS
individual treatment effects, personalized medicine, validation

1 | INTRODUCTION

The premise of personalized medicine is that individual differences in response to treatment are often large. It further supposes that outcomes can be improved by selecting, from among a set of possible treatments, a treatment with which an individual person is more likely to have a positive outcome.[1,2] While some examples of personalized medicine are very simple (lymphoma with the TP53 mutation responds poorly to chemotherapy[3] and so immunotherapies are indicated[4]), in other areas there is evidence for complex interactions between multiple factors in predicting differential treatment response. Alcohol use disorder (AUD) is an example of one condition that is characterized by complex etiology and heterogeneity in course, such that different treatments are very likely to be more or less effective for any one individual.[5-7] These situations call for new *methods for personalized medicine* which are characterized by the use of a model or algorithm to predict individual treatment response rather than identification of one moderating characteristic.[8,9] Given the complexity and heterogeneity that characterize AUD, we use treatment of AUD to frame discussion of our proposed method for personalized medicine.

One of the challenges common to many of these new methods is that predictions of differential treatment response are made for individuals, but each person's outcome is only ever observed under one treatment condition. Thus, the differential treatment response for an individual (the difference in outcomes under different treatment conditions) is not observed and it is both challenging to estimate heterogeneity in treatment response and to validate individual predictions. In this paper, we propose and test an approach for validating predictions across samples that uses matching in the test (validation) sample to match individuals in the treatment and control arms based on their predicted treatment response and their predicted response under control. We refer to validation as the process of showing that predictions from a training sample are replicated in a test sample. While the method could be used for internal cross-validation, evidence for validity is stronger when it can be shown that results replicate, or partially replicate, in independent samples.[10] The method we propose can provide evidence for validation through external replication or partial replication – where predictions from the training data are useful for determining the approximate individual treatment effect relative to other trial participants in a separate test dataset but where the magnitude of treatment response might be over or underestimated. Partial replication might be expected, for example, when the intervention has been slightly modified in the study providing the test data.

1.1 | Individual treatment effects for AUD

There is strong evidence for heterogeneity in the etiology, clinical symptoms, clinical trajectories, and effectiveness of particular treatments in AUD. Many individual factors including negative affective states, coping resources, comorbid psychiatric conditions and substance use, social functioning, cognitive functioning, self-efficacy, and motivation to change may contribute to AUD treatment response.[11] 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.[12] However, identifying patient characteristics that are associated with better outcomes in specific treatments in order to make clinical decisions about treatment has remained elusive.[13] We first describe an algorithm for validating individual predictions for a continuous outcome before we evaluate the approach in simulations. Validation of predictions from an AUD treatment study are subsequently tested before we end with a discussion.

2 | A VALIDATION APPROACH FOR INDIVIDUAL PREDICTIONS

An important question for any method designed to be implemented in clinical practice is whether initial results hold for new cases. This is especially important in personalized medicine where the result is a prediction about the effect of treatment vs control (or the comparison of two treatments) for an individual. Patients and health care providers want to know that the specific predictions made will hold outside of the initial clinical trial from which they were generated. To evaluate if this is the case and to gain confidence in the predictions, we wish to develop a strategy that seeks to confirm the validity of the prediction derived from a training dataset in an independent (test) dataset. We begin with an algorithm (Algorithm 1) which can be used for validation of predictions of continuous outcomes assessed at one time point. We then describe the Algorithm 2 which is used to obtain predicted individual treatment effects (PITEs) for any individual with baseline data using a training dataset from a randomized clinical trial. As the PITE approach utilizes predictions of outcomes under two treatment conditions, which are not both observed for a single patient, we then introduce Algorithm 3, which obtains pseudo-observations in a validation data set.

The process of validation is:

1. Obtain PITEs using data from a clinical trial and Algorithm 2 below.
2. Obtain data from a second (test) trial including the same covariates and outcome from the trial in step 1.
3. Use the algorithm fit with the training data to obtain PITEs in test data.
4. Use Algorithm 3 below to obtain pseudo-observations of individual treatment effects in the test data.
5. Use Algorithm 1 to evaluate whether the PITEs from the training data are replicated in the test data.

2.1 | Algorithm 1 – Validation of individual predictions with continuous outcomes

1. Train the predictive model to obtain individual predictions of some quantity, θ, on the training data.
2. Obtain individual predictions for all observations in the test data, $\widehat{\theta}_{test}$, using the model from training data.
3. Calculate the observed value of θ for all observations in the test data, θ_{obs}.
4. Fit $\theta_{obs} = \beta_0 + \beta_1 * \widehat{\theta}_{test}$.
5. Obtain the $1 - \alpha$ confidence interval for β_1 defined as $\left[\widehat{\beta}_1 \pm z_{1-\frac{\alpha}{2}} * SE\left(\widehat{\beta}_1\right)\right]$.
6. Conclude that the individual predictions are *compatible* with the new data if the confidence interval includes 1.

We note that the confidence interval from step 5 is correct when testing whether predictions in the training data are equal to those observed in the test data.

2.1.1 | Obtaining predicted individual treatment effects

In Algorithm 1 we have not made any explicit assumptions about the type of prediction that we are interested in and hence one can imagine the validation of a model that investigates the response to a single treatment. For this paper, however, we will focus on the question of differential treatment effect and use the PITE framework to define the estimand of interest and demonstrate our validation algorithm.[14-16] The PITE approach, which is based on potential outcomes, is similar to other personalized medicine methods[17] in that an array of baseline covariates are used to obtain predictions of outcomes under each treatment (note that Conditional Average Treatment Effects are very similar, differing in their application[18]). The PITE approach uses the following algorithm.

2.2 | Algorithm 2 – Obtain PITEs

1. Train a predictive model for the outcome for those in the control condition, $Y_i^c = f_c(x_i) + \varepsilon_{ic}$.
2. Train a predictive model for those in treatment, $Y_i^t = f_t(x_i) + \varepsilon_{it}$.
3. Compute the predicted value under control, $\widehat{Y}_i^c$, and under treatment, $\widehat{Y}_i^t$, for an individual using their values on baseline covariates.
4. Calculate the individual's PITE as the difference in these two predicted values: $PITE_i = \widehat{Y}_i^t - \widehat{Y}_i^c$.

The result of this algorithm are PITEs which can then be validated using Algorithm 1. A strength of the PITE approach is that any function which results in individual prediction could be used for $f_c(x_i)$, and $f_t(x_i)$.

2.2.1 | Estimate pseudo-observations of treatment effect for those in the test data

Because of the 'fundamental problem with causal inference', the true individual treatment effect for each individual (ITE_i) can never be directly observed.[19] We obtain a pseudo-observation[20] ($\widehat{ITE}_i$) of the treatment effect in the test (validation) sample by matching individuals in the treatment and control conditions based on their predicted value under treatment and predicted value under control, estimated using baseline covariates and with models from the training data using the following algorithm.

2.3 | Algorithm 3 – Obtain pseudo-observations of treatment effects in test data

1. Obtain a test data set in which individuals were randomized to treatment conditions and that contains baseline covariates used for estimating PITEs in the training data (Algorithm 2) as well as observed treatment outcomes.
2. Estimate $\widehat{Y}_i^c$ and $\widehat{Y}_i^t$ for each individual in the test data using the algorithm obtained from the training dataset.
3. Use a matching algorithm to match each individual in the treatment condition to an individual in control based on $\widehat{Y}_i^c$ and $\widehat{Y}_i^t$.

4. The pseudo-observed treatment effect for each individual in the test data, $\widehat{ITE}_i$, is calculated as the difference in the realized outcome Y for the treatment individual and their matched control individual.

Because $\widehat{Y}_i^c$ and $\widehat{Y}_i^t$ are derived from baseline covariates and an algorithm trained on the independent training dataset, $\widehat{ITE}_i$ and PITE$_\mathrm{i}$ are independent with the former being a function of the realized value for a matching pair of individuals in the test data and the latter being the prediction based on the training data. The $\widehat{ITE}_i$ is used as θ_{obs} in Algorithm 1 and the PITE$_\mathrm{i}$ is used as $\widehat{\theta}_{test}$, the proposed validation test can now be completed.

We note that the 1:1 matching procedure requires equal sample sizes in each treatment condition. One solution to this that has been proposed[21] is to draw repeated random samples equal to the size of the smaller condition, match and obtain $\widehat{ITE}_i$ within each sample, and then, for each person, average their $\widehat{ITE}_i$ across all random samples.

2.4 | Assumptions

To be considered causal, PITEs require the assumptions of consistency (the observed and potential outcomes are the same for the treatment assigned), exchangeability (potential outcomes are independent of treatment assignment given the observed characteristics), and positivity (every patient has a non-zero probability of receiving each treatment) which are typically met when using data from randomized trials with covariates assessed before randomization.[17] Additional assumptions needed to establish that PITEs are replicated in the test trial include: treatment and outcome stability (the treatment and outcome are the same in both studies and there is no spill-over effect between studies), equivalence of causal estimands, unbiased prediction of effects, and the effects in the training study are correctly reported with no mistakes.[10] Assumptions related to estimation of PITEs are straight forward, but assumptions required establish replication in a test trial are more onerous.

Some of these assumptions, such as outcome stability, clearly apply to all uses of our validation approach. Other assumptions, such as the equivalence of causal estimands, which includes that the populations are the equivalent especially with regards to any moderators of treatment effects, are meet in our approach so long as the moderators of importance are measured, included in the predictive models, and their effects are correctly specified. Our approach to validation adds the assumption that there are meaningful individual differences in treatment effects in training data. Lacking this, the expected value for β_1 in Algorithm 1 is not 1. The approach also requires adequate matches in the treatment and control conditions for the test data. While this assumption should be evaluated, given that treatment is randomized in both the training and test data, we expect few problems in practice.

Assumptions needed to establish complete replication of causal effects are quite daunting. From the perspective of clinical utility, assessing the procedure where partial replication is expected, or when out of sample individuals are similar, but not identical to the original sample, will often be of great interest. By looking at replication as conditions change, it is possible to begin to assess whether predictions apply outside of the, often restrictive, conditions in the original clinical trial. While assessing replication assumes that the populations, treatments, and covariates from each are comparable. When population, treatment, and/or covariates differ across training and test trials then the question to be addressed is whether and to what degree predictions from the initial training study hold.

Finally, Algorithm 1 assumes that the relationship between PITE predictions from the training data and estimated PITE observations in the test data is linear and confidence intervals assume that a parametric linear model is reasonable. Linearity will hold if predictions are perfectly replicated but otherwise should be examined and substituting a non-linear function into the algorithm could provide meaningful information about when results replicate.

2.5 | Related research

Independent of this study other researchers have recently developed methods for cross-validation of individual predictions of treatment effects. Gao, Hastie, and Tabshirani have proposed a novel matching method for obtaining pseudo-observations for treatment effects (this matches our algorithm 3), focusing on different approaches to matching and extending the approach to non-normal outcomes.[20] While Gao et al. supports the use of matching it does not directly address how to establish that a result is validated. In another paper on assessing performance of personalized approaches, Efthimiou and colleagues suggested the regression of the estimated individual treatment effect on the predicted effect, as in our approach, a regression weight of 1 indicating correct model calibration.[21] That paper focuses on internal validation and does not provide an evaluation of the performance of the method.

In contrast to previous research, we focus on the external validation of results from one randomized trial using independent data. This aim raises some unique issues which are the focus of our simulations. One issue is that results may be partially validated in a separate trial, for example, this may be the case when the trials differ in population, treatments, or covariates. It will often be very useful to know whether, and to what extent, results from one trial hold in a test trial in the presence of differences in design across the trials. A second issue concerns the presence of nuisance variables—variables included in the predictive model which do not predict treatment response. We examine the performance of the approach as a function of sample size, nuisance variables, and partial replication. We also provide detailed steps for how to implement this method for external validation.

3 | SIMULATION STUDY

3.1 | Aims

This simulation study is primarily concerned with establishing a method for providing external validation of predictions of individual treatment effects. We aim to offer proof-of-concept of the approach, specified above, for external validation and show the conditions under which the method is viable or fallible. This simulation study aims, primarily, to investigate how robust the method is under alternative conditions.

3.2 | Generating training and test data

The simulated training datasets included 7 variables that were responsible for individual differences in the effects of the treatment in the training data. We chose 7 variables to replicate previous methodological work on PITEs.[22] Of the 7 covariates that predicted treatment response, 6 were drawn from a standard normal distribution and the last from a binomial distribution with a probability of .5. The outcome was generated using these 7 covariates with separate regression weights under control (.09, −.20, .10, .18, −.20, −.10, and −.09) and treatment (−.09, .20, −.10, −.18, .20, .20, and .09), and with a random error term with a SD (SD) of 1. The effects were mostly centered around 0 so that there was no average main effect of each covariate, corresponding with what was observed in the trial data used for the applied example, and for simplicity, the intercept was assumed to be 0 for both groups, indicating no average treatment effect. We conducted a sensitivity analysis where the regression weights were not centered on zero (see Supplemental Material) and found no meaningfully different results.

The vectors of effects were chosen, after a process of systematic modification to ensure consistency, so that the effect size of the PITE (SD of the PITE, divided by the SD of the outcome under control) was .75 (similar to the total estimated value for the trial data in our applied example) for a sample size of 600. While limited literature with individual differences in treatment effects is available, this is most likely a moderately large effect size. A sensitivity analysis reported in the Supplementary Materials found similar results with larger confidence intervals for an effect of .375. In each case test data for validation was generated using the parameter estimates observed in the training data, rather than duplicating the parameters used to generate the training data. This is because initial simulations found that while the validation parameter (β_1 in algorithm 1) remains the same under either condition, there is additional variability in the parameter estimates in the second approach which results in confidence intervals being somewhat too small. It is thus important to note that the confidence intervals from algorithm 1 are evaluated with respect to whether the effect in the training data is replicated in the test data.

3.3 | Simulation study target

This approach is primarily concerned with testing whether the observed heterogeneity in treatment effects is replicated in independent test data. Following generation of the test data we used the process in Algorithm 3 to obtain pseudo-observed PITEs for each individual in the test data: (1) the predicted outcome under control and predicted outcome under treatment for each individual was obtained using the training algorithm; (2) individuals in the treatment arm were matched with individuals in the control arm based on these two values using the MatchIt package[23] in R with nearest neighbor matching using Mahalanobis distance; (3) Within each matched pair the observed outcome for the control individual was subtracted from the outcome for the treatment individual to obtain the pseudo-observed PITE for the pair; (4) the pseudo-observed PITE was on the predicted PITE across the entire test sample; and (5) the regression weight (validation parameter) and confidence interval of the resulting slope are the parameters of interest. The validation parameter assesses the extent to which the predictions in the training data are replicated in the test data.

3.4 | Simulation conditions

To examine the proposed validation approach we conducted simulations which under 4 different conditions: (1) sample size (with 4 different levels including 300; 600; 1200; and 12 000, 50% in treatment and 50% in control groups, in both training and test datasets); (2) whether results in the training data were replicated in the test data (3-levels included perfect replication; partial replication; and no replication); (3) predictive model (linear regression and LASSO); and (4) nuisance variables (3-levels including none; 10; and 20). The first 3 are fully crossed, nuisance variables are only run for the perfect replication condition as they are expected to lead to bias in the validation parameter which is best shown in this condition. Thus, there are 80 different simulation conditions. Each simulation setting was run with 1000 training datasets which were then validated in 1000 test datasets.

The sample sizes chosen were intended to show performance of the validation approach for moderate to large clinical trials for which this method is appropriate, the 12 000 conditions are included to show performance at the limit. The replication condition is included to show that the validation parameter performs as expected when there is no replication and complete replication as well as when results only partially replicate. The different levels of replication were achieved by multiplying the coefficients used to generate the test data by 0 (results were not replicated); .6 for partial replication (which reduced the target effect size of the PITE by half with a sample of 600); and 1 for full replication. The partial replication condition results in the individual differences in treatment effects being reduced in the test data resulting in the average differential effect being reduced. This might be caused, for example, by the treatment in the test sample being somewhat adapted from the treatment in the training sample. We expect to see partial validation indicated by a regression weight between 0 and 1. We used two predictive models to estimate PITEs – the linear model and LASSO regression (implemented via the lars package[24]) in steps 1 and 2 of Algorithm 2. In conditions using LASSO, the penalty parameter (lambda) was chosen using 10-fold cross-validation. The last factor varied across simulations is the presence of nuisance variables, which are baseline covariates included in the predictive model although in practice they do not predict treatment response. We expect that most PITE applications will include some variables like this because investigators do not want to miss true heterogeneity. We anticipate that when nuisance variables are present, LASSO will result in less noise in the predictions and hence better validation performance.

3.5 | Performance measures

Efthimiou and colleagues propose assessing calibration of predictions by regressing the observed outcome on the predicted outcome with a regression weight of 1 indicating perfect calibration.[21] We use this validation parameter to assess whether results from a training sample are replicated in a test sample, we test bias by comparing the average of this coefficient for each simulation condition to 0 (for no replication) and 1 for replication. Performance of confidence intervals is the proportion of simulations under each condition in which the confidence interval for the validation parameter includes 0 or 1 (depending on condition). Further performance measures considered included the extent to which LASSO over-selected nuisance variables and under-selected true covariates in conditions with nuisance variables.

3.6 | Results

3.6.1 | Assessing perfect replication

When the test data perfectly replicated the training data we see no meaningful bias in the validation parameter (see the slope in Table 1). Coverage rates for the slope (95% CI included 1) and intercept (95% CI included 0) across all simulations

TABLE 1 Performance of validation approach across simulation conditions.

		Linear model						LASSO					
		Slope			Intercept			Slope			Intercept		
	N	**Mean**	**SD**	**Coverage (of 1)**[a]	**Mean**	**SD**	**Coverage (of 0)**	**Mean**	**SD**	**Coverage (of 1)**[a]	**Mean**	**SD**	**Coverage (of 0)**
Training effects multiplied by 1	300	0.997	0.148	95.6%	0.000	0.120	95.3%	0.993	0.199	95.9%	0.019	0.136	91.2%
	600	0.993	0.109	93.8%	−0.002	0.084	95.0%	0.992	0.118	93.7%	−0.002	0.084	95.0%
	1200	1.002	0.075	94.7%	−0.002	0.059	94.7%	1.002	0.077	94.6%	0.003	0.060	94.2%
	12 000	1.000	0.023	95.7%	−0.001	0.018	95.9%	1.000	0.023	95.7%	−0.001	0.018	95.9%
Training effects multiplied by 0.6	300	0.597	0.148	NA	−0.002	0.140	90.2%	0.593	0.199	NA	0.010	0.152	86.0%
	600	0.593	0.109	NA	−0.002	0.098	95.2%	0.593	0.118	NA	0.004	0.100	95.1%
	1200	0.602	0.075	NA	−0.002	0.068	90.8%	0.602	0.077	NA	0.000	0.068	90.5%
	12 000	0.600	0.023	NA	0.000	0.020	91.9%	0.600	0.023	NA	0.000	0.020	91.9%
Training effects multiplied by 0	300	−0.003	0.148	95.4%	−0.005	0.210	72.8%	−0.007	0.200	95.3%	−0.005	0.212	72.5%
	600	−0.006	0.109	93.9%	−0.003	0.148	86.4%	−0.007	0.118	93.8%	−0.003	0.148	86.5%
	1200	0.002	0.075	94.9%	−0.003	0.102	76.1%	0.002	0.077	94.5%	−0.003	0.102	76.0%
	12 000	0.000	0.023	95.8%	0.000	0.030	77.1%	0.000	0.023	95.8%	0.000	0.030	77.1%
Training effects multiplied by 1 w/ 10 nuisance variables	300	0.824	0.154	72.0%	−0.012	0.174	84.0%	0.890	0.276	89.0%	−0.007	0.142	91.7%
	600	0.905	0.112	81.6%	−0.009	0.117	82.4%	0.948	0.150	91.0%	−0.006	0.096	92.1%
	1200	0.948	0.076	88.7%	−0.004	0.085	82.8%	0.971	0.088	94.3%	−0.002	0.072	89.1%
	12 000	0.996	0.024	94.1%	−0.001	0.025	83.8%	0.998	0.025	93.2%	−0.001	0.022	88.6%
Training effects multiplied by 1 w/ 20 nuisance variables	300	0.679	0.162	32.2%	−0.030	0.222	71.5%	0.803	0.346	84.1%	−0.021	0.149	90.3%
	600	0.823	0.111	55.3%	−0.021	0.143	75.4%	0.925	0.158	89.9%	−0.008	0.102	89.5%
	1200	0.906	0.076	73.4%	−0.008	0.106	71.2%	0.963	0.091	92.5%	−0.004	0.076	87.2%
	12 000	0.989	0.024	92.1%	0.002	0.031	75.1%	0.996	0.025	94.9%	0.001	0.023	89.4%

[a] Coverage for the slope when PITEs are not replicated is assessed around 0.

neared 95%. Results for the 100% replication condition looked similar when using LASSO to generate predicted PITEs. For large sample sizes ($N = 12\,000$ and $N = 1200$) LASSO typically selected all 7 important variables (99.8% and 71.8% of the time, respectively). For smaller samples the probability that one ($N = 300$, 29.8% and $N = 600$, 31.5%) or two ($N = 300$, 19.4% and $N = 600$, 12.3%) variables underlying heterogeneity were not selected increased.

3.6.2 | Assessing partial replication

We next examined partial replication wherein the coefficients used to generate the test data were multiplied by 0.6 (reducing the SD of PITEs by half) from that observed in the training data. Using coefficients from the linear model to calculate predicted PITEs in the test data, we observed that the slope of the observed PITE on predicted PITE was approximately 0.6. In all cases the confidence intervals for the slopes excluded the conclusion that there was no replication (slope = 0) or that there was perfect replication (slope = 1). We observed very similar patterns among the models using penalized regressions in the partial replication conditions. We note that for the partial replication and no replication conditions the means of the intercept were close to 0 (minimal bias observed) but the coverage rates were low in many cases. We attribute this to the lack of independence between slopes and intercepts (when the slope is high or low that will increase variance in the intercept). The intercepts reflect estimates of mean differences between treatment and control, which is not the focus of this work and which are typically assessed independently of PITEs (ie, by assessing mean differences between groups). Thus, this is not a major concern for our application.

TABLE 2 LASSO variable selection across simulation conditions.

		Number of covariates selected (tx)				Number of covariates selected (c)				% nuisance variables selected (tx)		% nuisance variables selected (c)	
	Condition	≤4	5	6	7	≤4	5	6	7	Mean	SD	Mean	SD
Training effects multiplied by 1 w/ 0 nuisance variables	$N = 300$	17.4%	18.1%	32.8%	31.7%	26.6%	20.6%	26.7%	26.1%	NA	NA	NA	NA
	$N = 600$	1.7%	10.8%	31.4%	56.1%	5.9%	13.7%	31.6%	48.8%	NA	NA	NA	NA
	$N = 1200$	0.1%	2.8%	24.8%	71.3%	0.5%	3.7%	24.5%	71.3%	NA	NA	NA	NA
	$N = 12\,000$	0.0%	0.0%	0.2%	99.8%	0.0%	0.0%	0.2%	99.8%	NA	NA	NA	NA
Training effects multiplied by 1 w/ 10 nuisance variables	$N = 300$	30.7%	27.9%	25.0%	16.4%	30.4%	21.7%	20.4%	10.4%	65%	24%	70%	24%
	$N = 600$	7.1%	18.3%	39.0%	35.6%	12.9%	22.1%	34.5%	30.5%	56%	22%	58%	24%
	$N = 1200$	0.5%	7.5%	35.7%	56.3%	1.4%	8.5%	36.1%	54.0%	53%	21%	55%	21%
	$N = 12\,000$	0.0%	0.0%	0.0%	100.0%	0.0%	0.0%	0.5%	99.5%	45%	21%	46%	21%
Training effects multiplied by 1 w/ 20 nuisance variables	$N = 300$	46.1%	25.4%	20.2%	8.4%	61.3%	20.1%	13.3%	5.2%	74%	19%	78%	19%
	$N = 600$	10.5%	25.0%	37.9%	26.6%	20.2%	26.7%	34.3%	18.8%	65%	18%	69%	18%
	$N = 1200$	0.9%	9.3%	41.4%	48.4%	20.0%	10.8%	40.3%	46.7%	61%	17%	63%	18%
	$N = 12\,000$	0.0%	0.0%	0.5%	99.5%	0.0%	0.0%	0.5%	99.5%	57%	17%	57%	17%

3.6.3 | Assessing a lack of replication

No replication between the training dataset and the test dataset was assessed by multiplying the coefficients used to generate the test data by 0. For both the linear model and LASSO, average slopes were very close to 0 across all conditions and the coverage rates of the 95% CIs for the slopes were very close to the expected proportion of .95.

3.6.4 | Incorporating nuisance variables

To examine the relative influence of noise on our validation procedure we included 10 and 20 nuisance variables. The linear model performed relatively poorly with average slopes as low as 0.7 with a sample size of 300 and reaching 1 only with a sample size of 12 000. Although LASSO did not completely correct for the inclusion of nuisance variables, it did improve upon linear model results with slopes approaching 1 with sample sizes of 600 or greater. Underestimation of slopes and lower coverage rates are much less severe with LASSO. Although LASSO effectively improved validation results in the presence of nuisance variables, it comes at the potential cost that important variables are excluded from the original PITEs and thus differences between people in treatment effects due to those variables is missed. Table 2 documents the number of important variables excluded and the number of nuisance variables included using LASSO with the default tuning parameters.

4 | APPLICATION TO TREATMENT OF ALCOHOL USE DISORDER

To demonstrate the use of this validation procedure in the context of predicting individual treatment effects in the treatment of AUD, we apply our validation procedure using data from COMBINE.[25] We use an example where the goal is to establish whether individual differences in the effects of a pharmacological intervention hold after a behavioral intervention is added, this is an example that frequently occurs in practice (individuals receive multiple treatments at once), where the proposed validation approach can be used to assess whether predictions under one treatment hold when the treatment is expanded to include multiple components, a scenario where we expect partial replication. Two research questions drove this application: (1) Do we observe individual heterogeneity in the effects of naltrexone compared to placebo on percent drinking days due to variation in baseline biomarkers and (2) if we find heterogeneity in treatment effects, does that finding hold in the presence of a Cognitive Behavioral Intervention (CBI)? To further personalized medicine approaches in the treatment of AUD, scholars have called for a greater investigation of biomarkers and their potential to predict individual differences in treatment response.[26,27] Although most of the research in this area focuses on the effects of genetic variation (single nucleotide polymorphisms [SNPs]), in this application, we use non-genetic biomarker data to generate PITEs with percent drinking days at the end of the trial treatment period (16 weeks after initiation of treatment) as the outcome.

4.1 | Methods for COMBINE application

The Combining Medications and Behavioral Interventions (COMBINE) trial is a randomized controlled trial that tested a broad range of treatment combinations for AUD in 1383 patients. Between January 2001 and January 2004 individuals diagnosed with AUD were recruited from 11 outpatient alcohol treatment clinics in academic centers across the United States and were assigned to one of nine different treatment conditions (4 pharmacological conditions: placebo, acamprosate, naltrexone, acamprosate and naltrexone; crossed with 2 behavioral conditions: cognitive behavioral intervention (CBI) and treatment as usual with the addition of a treatment as usual with no medicine condition). One of the primary outcomes in the original trial, which we use for this application, was percentage of drinking days in the previous month. Data from COMBINE present an opportunity for examining heterogeneity in treatment response among AUD treatment-seekers and for using the validation approach proposed to test whether individual predictions for the effects of a pharmacological treatment (Naltrexone) hold when the treated individuals are also receiving a second (behavioral) intervention at the same time.

For this example, to examine heterogeneity in treatment effects, we compare two arms: naltrexone and medication management ($N = 154$) vs placebo and medication management ($N = 153$) for the initial training of PITEs. To test the validation approach, we assess whether individual differences observed are also replicated in a test dataset comprised of the naltrexone and CBI ($N = 155$) vs the placebo and CBI ($N = 156$) arms of the trial, for a total of 618 individuals in training and test samples.

4.2 | Predictors of individual differences

Biomarkers were selected from the laboratory assessments conducted as part of the full COMBINE assessment battery. First, we excluded all variables ($N = 90$) with over 95% missingness. We standardized all baseline biomarkers and examined outliers. We performed a 90% winsorization to accommodate outliers, meaning that for those cases with values below the 5th percentile or above the 95th percentile of the distribution for a given variable, the outlier was replaced with the 5th or 95th percentile respectively. Next, we examined multicollinearity among predictors by performing two linear regressions—one among treated individuals, and another among controls with each of the remaining baseline biomarkers predicting percent days drinking in the last month of treatment. For each regression, we examined variance inflation factors (VIFs) for all the predictors. We eliminated predictors with the highest VIF, one at a time, until all VIFs were less than 5.[28] We were left with 27 baseline biomarkers to predict heterogeneity in treatment effects: magnesium (mg/DL), sodium (mEq/L), calcium (mg/dL), potassium (mEq/L), phosphorus (mg/dL), bicarbonate (mEq/L), creatinine (mg/dL), blood urea nitrogen (mg/dL), glucose (mg/dL), uric acid (mg/dL), alanine transaminase (IU/L), total bilirubin (mg/dL), alkaline phosphatase (IU/L), lactate dehydrogenase (IU/L), total protein (g/dL), albumin (g/dL), hemoglobin (g/dL), hematocrit (percent), white blood cells ($\times10^3$/uL), lymphocytes (percent), monocytes (percent), eosinophils (percent), basophils (percent), mean corpuscular hemoglobin, mean corpuscular hemoglobin concentration (gHb/dL), urine pH, and urine specific gravity.

4.3 | Analysis for COMBINE application

We first computed PITEs in the training data set (medication management only) and performed a permutation test to determine whether the 27 biomarkers predicted significant differences in PITEs.[22] The null hypothesis of the permutation was that there are no individual differences in treatment effects on the basis of the covariates specified. To test the null hypothesis, the permutation test randomly reassigns treatment condition 1000 times, calculates PITEs for each permutation and determines whether the SD of the PITEs observed from the original data is greater than the SD of the PITEs from 95% of the datasets with randomly permuted treatment assignments (significance threshold of 0.05). A higher observed SD of the PITEs, compared to the SDs in the datasets with randomly permuted treatment assignment is interpreted as evidence that the observed heterogeneity is more than would be expected due to random chance. We then apply our validation procedure to a test dataset (medication management + CBI) that employed comparable treatment and control conditions (naltrexone vs placebo) to determine if results from the training data would replicate in a new sample which had the same intervention plus an additional psychosocial treatment. Due to different scales of the baseline covariates and the outcome, all continuous variables were standardized.

4.4 | Results for COMBINE application

PITEs in the training data, calculated following Algorithm 2 using linear models, demonstrated a positive skew and a SD of 0.776. The results of the permutation test indicate that this SD is not significantly greater than expected due to chance ($P = 0.198$), suggesting that we cannot be confident that observed individual differences in treatment effects as a function of these biomarkers are greater than chance. Despite lacking confidence in the existence of heterogeneity in treatment effects due to the selected biomarker data, we proceeded with the validation procedure to determine whether we would find comparable results in the test data. The validation procedure provides another approach for establishing whether individual differences observed are reliable.

We calculated PITEs for the test data using the algorithm from the training data and the test data covariate matrix, matched treated and control individuals, obtained pseudo-observed PITEs for the pair, and regressed these on predicted PITEs. Results from this regression indicate that the slope of the predicted PITE was close to zero with a confidence interval which indicated a great deal of uncertainty around zero (B = −0.100, SE = 0.169, 95% CI: [−0.436, 0.235]). We interpret the results as secondary support for the result that PITEs in the training dataset are not greater than chance.

5 | DISCUSSION

This paper presents a generic algorithm for validation (replication) of predictions of treatment effects with continuous outcomes from one sample in a new sample. This is a difficult problem because treatment response is not observed for an individual, but it is a problem for which an answer is needed if individual predictions are to be used in clinical practice. We illustrate it in the context of differential treatment effects, show its use in an example of treatment of alcohol use disorder, and verify its statistical properties through simulations. We find that the validation algorithm correctly confirms validation in cases of full replication and lack of validation, but also provides insights into cases of partial validation. An important finding is that the presence of nuisance variables detrimentally impacts algorithm performance leading to downward bias in the validation parameter which can be partially reduced though the use of variable selection methods. An advantage of the approach is that it can be widely applied to different predictive methods. We note that meaningful validation presumes the presence of meaningful treatment response heterogeneity in the training data and that when heterogeneity in the training data does not meet traditional measures of statistical significance this test provides secondary evidence which may be useful.

The use of PITEs for the treatment of alcohol use disorder, as well as other substance use and mental health diagnoses, requires external validation of individual predictions. Our method, which is similar to an approach described by Efthimiou and colleagues,[21] provides a way to do this with independent validation data. A primary limitation of this method is its applicability only to continuous outcomes at a discrete timepoint (although inclusion of the outcome assessed pre-randomization allows for simple pre-post analysis). Future efforts will include expanding the approach to accommodate binary outcomes and more complex longitudinal models. Other limitations of the method are that nuisance variables reduce values of the validation parameter and sensitivity to small sample sizes (especially with regards to distinguishing nuisance variables from partial replication). Conversely, large sample sizes might drive down the size of P-values, thus increasing the risk of Type 1 error. However, the largest sample sizes that we used in this simulation study as a demonstration would be unrealistic to expect in most clinical trial datasets. Further, while we did not test whether this approach can identify small groups of individuals who differ substantially from the average treatment effect, we do not believe that it is optimal in that situation. Finally, we emphasize that PITEs assume that the average treatment effects (conditional on covariates) are properly estimated, if not, the predictions will be systematically biased. It is important to note that confidence intervals are for whether initial estimates are replicated in test data rather than for whether underlying true estimates are the same in two samples. Notwithstanding the need for these additional issues to be addressed, this approach is important for the translation of personalized medicine predictions to clinical practice.

FUNDING INFORMATION

This paper was partially supported by grants from NIAAA: R01AA030264 (PI: M. Lee Van Horn), T32AA018108 (PI: Katie Witkiewitz), and R01AA022328 (PI: Katie Witkiewitz).

CONFLICT OF INTEREST STATEMENT

The authors declare no conflicts of interest.

DATA AVAILABILITY STATEMENT

The data that support the findings of this study are available from National Institute on Alcohol Abuse and Alcoholism. Restrictions apply to the availability of these data, which were used under license for this study. Data are available from http://www.niaaa.nih.gov with the permission of National Institute on Alcohol Abuse and Alcoholism.

ORCID

Alena Kuhlemeier https://orcid.org/0000-0002-1917-3230
Thomas Jaki https://orcid.org/0000-0002-1096-188X

REFERENCES

1. Kessler RC, Luedtke A. Pragmatic precision psychiatry—A new direction for optimizing treatment selection. *JAMA Psychiatry*. 2021;Online bef: 1–7.;78:1384-1390. doi:10.1001/jamapsychiatry.2021.2500
2. Witkiewitz K, Roos CR, Mann K, Kranzler HR. Advancing precision medicine for Alcohol use disorder: replication and extension of reward drinking as a predictor of naltrexone response. *Alcohol Clin Exp Res*. 2019;43(11):2395-2405. doi:10.1111/acer.14183
3. Maddocks K. Update on mantle cell lymphoma. *Blood*. 2018;132(16):1647-1656.
4. Cortelazzo S, Ponzoni M, Ferreri AJM, Dreyling M. Mantle cell lymphoma. *Crit Rev Oncol Hematol*. 2020;153:103038.
5. Litten RZ, Ryan ML, Falk DE, Reilly M, Fertig JB, Koob GF. Heterogeneity of alcohol use disorder: understanding mechanisms to advance personalized treatment. *Alcohol Clin Exp Res*. 2015;39(4):579-584. doi:10.1111/acer.12669
6. Boness CL, Witkiewitz K. Precision medicine in alcohol use disorder: mapping etiologic and maintenance mechanisms to mechanisms of behavior change to improve patient outcomes. *Exp Clin Psychopharmacol*. 2022;31:769-779. doi:10.1037/pha0000613
7. Witkiewitz K, Litten RZ, Leggio L. Advances in the science and treatment of alcohol use disorder. *Sci Adv*. 2019;5:eaax4043.
8. Cirillo D, Valencia A. Big data analytics for personalized medicine. *Curr Opin Biotechnol*. 2019;58:161-167. doi:10.1016/j.copbio.2019.03.004
9. Malenica I, Phillips RV, Chambaz A, Hubbard AE, Pirracchio R, van der Laan MJ. Personalized online ensemble machine learning with applications for dynamic data streams. *Stat Med*. 2023;42:1013-1044. doi:10.1002/sim.9655
10. Steiner PM, Wong VC, Anglin K. A causal replication framework for designing and assessing replication efforts. *Z Psychol*. 2019;227(4):280-292.
11. Witkiewitz K, Marlatt GA. Relapse prevention for alcohol and drug problems: that was Zen, this is Tao. *Am Psychol*. 2004;59(4):224-235. doi:10.1037/0003-066X.59.4.224
12. Sliedrecht W, de Waart R, Witkiewitz K, Roozen HG. Alcohol use disorder relapse factors: a systematic review. *Psychiatry Res*. 2019;278:97-115. doi:10.1016/j.psychres.2019.05.038
13. Mann K, Roos CR, Hoffmann S, et al. Precision medicine in alcohol dependence: a controlled trial testing pharmacotherapy response among reward and relief drinking phenotypes. *Neuropsychopharmacology*. 2018;43(4):891-899. doi:10.1038/npp.2017.282
14. Ballarini NM, Rosenkranz GK, Jaki T, Konig F, Posch M. Subgroup identification in clinical trials via the predicted individual treatment effect. *PLoS One*. 2018;13(10):1-22. doi:10.1371/journal.pone.0205971
15. Kuhlemeier A, Desai Y, Tonigan AA, et al. Applying methods for personalized medicine to the treatment of alcohol use disorder. *J Consult Clin Psychol*. 2021;89(4):288-300.
16. Lamont A, Lyons MD, Jaki T, et al. Identification of predicted individual treatment effects in randomized clinical trials. *Stat Methods Med Res*. 2018;27(1):142-157. doi:10.1177/0962280215623981
17. Hoogland J, IntHout J, Belias M, et al. A tutorial on individualized treatment effect prediction from randomized trials with binary endpoint. *Stat Med*. 2021;40(26):5961-5981.
18. Vegetabile BG. On the distinction between "conditional average treatment effects (CATE)" and "individual treatment effects (ITE)" under ignorability assumptions. *arXiv preprint arXiv:210804939*. 2021.
19. Holland PW. Statistics and causal inference. *J Am Stat Assoc*. 1986;81(396):945-960.
20. Gao Z, Hastie T, Tibshirani R. Assessment of heterogeneous treatment effect estimation accuracy via matching. *Stat Med*. 2021;40:3990-4013.

21. Efthimiou O, Hoogland J, Debray TPA, et al. Measuring the performance of prediction models to personalized treatment choice. *Stat Med*. 2023;42(8):1188-1206.
22. Chang C, Jaki T, Sadiq MS, et al. A permutation test for assessing the presence of individual differences in treatment effects. *Stat Methods Med Res*. 2021;30(11):2369-2381. doi:10.1177/09622802211033640
23. Ho DE, Imai K, King G, Stuart EA. MatchIt: nonparametric preprocessing for parametric causal inference. *J Stat Softw*. 2011;42(8):1-28. doi:10.18637/jss.v042.i08
24. Efron B, Hastie T, Johnstone I, Tibshirani R. Least angle regression. *Ann Stat*. 2004;32(2):407-499.
25. Anton RF, Malley SSO, Ciraulo DA, et al. Combined pharmacotherapies and behavioral interventions for alcoholism (COMBINE trial). *JAMA J Am Med Assoc*. 2006;295(17):2003-2017.
26. Ray LA, Grodin EN, Leggio L, et al. The future of translational research on alcohol use disorder. *Addict Biol*. 2021;26(2):e12903. doi:10.1111/adb.12903
27. Heilig M, Sommer WH, Spanagel R. The need for treatment responsive translational biomarkers in alcoholism research. *Curr Top Behav Neurosci*. 2016;28:151-171. doi:10.1007/7854_2015_5006
28. Stine RA. Graphical interpretation of variance inflation factors. *Am Stat*. 1995;49(1):53-56.

SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

How to cite this article: Kuhlemeier A, Jaki T, Witkiewitz K, Stuart EA, Van Horn ML. Validation of predicted individual treatment effects in out of sample respondents. *Statistics in Medicine*. 2024;43(22):4349-4360. doi: 10.1002/sim.10187