


ORIGINAL ARTICLE

Evaluating the heterogeneous effect of a modifiable risk factor on suicide: The case of vitamin D deficiency

Jose R. Zubizarreta^{1,2,3} | John C. Umhau⁴ | Patricia A. Deuster⁵ |
Lisa A. Brenner^{6,7} | Andrew J. King¹ | Maria V. Petukhova¹ | Nancy A. Sampson¹ |
Boris Tizenberg¹⁰ | Sanjaya K. Upadhyaya¹⁰ | Jill A. RachBeisel⁸ |
Elizabeth A. Streeten⁹ | Ronald C. Kessler¹  | Teodor T. Postolache^{7,10,11}

¹Department of Health Care Policy, Harvard Medical School, Boston, Massachusetts, USA

²Department of Statistics, Harvard University, Cambridge, Massachusetts, USA

³Department of Biostatistics, Harvard Chan School of Public Health, Boston, Massachusetts, USA

⁴Alcohol Recovery Medicine, Potomac, Maryland, USA

⁵Consortium for Health and Military Performance, Department of Military & Emergency Medicine, F. Edward Hébert School of Medicine, Uniformed Services University, Bethesda, Maryland, USA

⁶University of Colorado Anschutz School of Medicine, Aurora, Colorado, USA

⁷VA Rocky Mountain Mental Illness Research Education and Clinical Center (MIRECC), Aurora, Colorado, USA

⁸Department of Psychiatry, University of Maryland School of Medicine, Baltimore, Maryland, USA

⁹Genetics and Personalized Medicine Clinic, Division of Endocrinology, Diabetes and Nutrition, University of Maryland School of Medicine, Baltimore, Maryland, USA

¹⁰Mood and Anxiety Program, Department of Psychiatry, University of Maryland School of Medicine, Baltimore, Maryland, USA

¹¹VISN 5 Capitol Health Care Network Mental Illness Research Education and Clinical Center (MIRECC), Baltimore, Maryland, USA

Correspondence

Ronald C. Kessler, Department of Health Care Policy, Harvard Medical School, 180 Longwood Ave, Ste 215, Boston, Massachusetts 02115-5899, USA.
Email: kessler@hcp.med.harvard.edu

Funding information

Patient-Centered Outcomes Research Institute, Grant/Award Number: ME-2019C1-16172; Rocky Mountain MIRECC for Suicide Prevention; Defense Advanced Research Projects Agency; Division of Intramural Research, National Institute of Allergy and Infectious Diseases

Abstract

Objectives: To illustrate the use of machine learning methods to search for heterogeneous effects of a target modifiable risk factor on suicide in observational studies. The illustration focuses on secondary analysis of a matched case-control study of vitamin D deficiency predicting subsequent suicide.

Methods: We describe a variety of machine learning methods to search for *prescriptive predictors*; that is, predictors of significant variation in the association between a target risk factor and subsequent suicide. In each case, the purpose is to evaluate the potential value of selective intervention on the target risk factor to prevent the outcome based on the provisional assumption that the target risk factor is causal. The approaches illustrated include risk modeling based on the super learner ensemble machine learning method, Least Absolute Shrinkage and Selection Operator (Lasso) penalized regression, and the causal forest algorithm.

Results: The logic of estimating heterogeneous intervention effects is explicated along with the illustration of some widely used methods for implementing this logic.

This is an open access article under the terms of the Creative Commons Attribution-NonCommercial-NoDerivs License, which permits use and distribution in any medium, provided the original work is properly cited, the use is non-commercial and no modifications or adaptations are made.

© 2021 The Authors. International Journal of Methods in Psychiatric Research published by John Wiley & Sons Ltd.

Conclusions: In addition to describing best practices in using the machine learning methods considered here, we close with a discussion of broader design and analysis issues in planning an observational study to investigate heterogeneous effects of a modifiable risk factor.

KEYWORDS

causal forest algorithm, heterogeneity of treatment effects (HTE), lasso penalized regression, precision medicine, prescriptive predictors, suicide, super learner

1 | INTRODUCTION

In recent years, there has been a growing number of studies aimed at predicting future health outcomes based on machine learning analyses applied to electronic medical records. A substantial literature of this sort has grown up, for example, to predict suicides and suicide attempts. Several reviews exist regarding this extensive literature (Reale et al., 2021; Rossom et al., 2021) along with critiques (Belsher et al., 2019; Bossarte et al., 2021; Kessler et al., 2020).

One of the most important of critiques is that models designed to improve allocation of expensive preventive interventions for suicide prediction need more specificity on the most appropriate target populations for such interventions and the ideal risk horizons for prediction (Kessler et al., 2019). The lack of specificity in suicide risk models is critical because a model developed to determine the appropriateness of, for example, involuntary hospitalization because of imminent suicide risk would be based on quite a different target population (i.e., patients presenting at an emergency room) and risk horizon (e.g., suicide over the next 48 h) than a model developed to target patients for outpatient suicide-focused psychotherapy. Yet many machine learning studies of suicide risk prediction, albeit with notable exceptions (e.g., Reale et al., 2021; Rossom et al., 2021), present nothing more than models applied to total populations over diverse risk periods with a lack of clarity about either the types of interventions the models are designed to target or the appropriate risk horizons (Barak-Corren et al., 2017; Gradus et al., 2020).

A related issue is that machine learning studies designed to target specific suicide prevention interventions focus largely on *high suicide risk* in the target population. It is important to realize, though, that patients at highest suicide risk might not be the ones most appropriate for a given intervention. A good example is the VA *Recovery Engagement And Coordination for Health-Veterans Enhanced Treatment* (REACH VET) Initiative (VA Office of Public and Intergovernmental Affairs, 2017). This initiative uses a machine learning model to target outreach case management interventions to the 0.1% of users of the Veterans Health Administration system with highest predicted suicide risk. Would this type of intervention be expected to prevent suicides among these extremely high-risk patients, the great majority of whom are already well-known to the treatment system? Or might this type of intervention be more successful in preventing suicides if it was directed to patients with somewhat lower, but still

elevated, risk? We have no way of answering this question from the design of the REACH VET implementation.

Research designed to address such questions is important because different interventions are likely to be optimal for different patient populations. Research that investigates this *heterogeneity of treatment effects* (HTE) is typically referred to as “precision medicine” research (Fernandes et al., 2017). This is an active area of investigation in psychiatry (Salazar de Pablo et al., 2021), but not yet in research on interventions for suicide prevention. Absence of suicide-focused HTE research is important because, aside from a few widely accepted universal interventions (Brodsky et al., 2018), suicide-focused interventions generally have relatively weak aggregate effects (Zalsman et al., 2016). This has limited the widespread dissemination of suicide prevention interventions. Indeed, some critics have gone so far as to suggest that suicide prevention research should be abandoned based on the weak intervention effects (Hoge, 2019). However, if these weak aggregate intervention effects reflect the existence of HTE, a case could be made for carrying out analyses to discover this heterogeneity and to implement different interventions with different subsets of patients.

HTE is likely in suicide-focused preventive interventions for two reasons. First, suicides occur in conjunction with manifold mental health disorders. Despite many similar issues in managing suicidal thoughts and behaviors across these disorders, meaningful differences are likely in causal risk factors across these disorders, leading to HTE. Second, suicide-focused interventions target intermediate outcomes, such as increased perceptions of belongingness in caring text interventions (Comtois et al., 2019) or “suicide drivers” in CAMS (*Collaborative Assessment and Management of Suicidality*) therapy (Jobes, 2012), each of which is an issue for only a subset of patients. This inevitably leads to weak aggregate effects, even if large effects exist among the subset of patients for whom the intervention focus is relevant. It is consequently of considerable importance to investigate the subset of patients for whom any given suicide-focused intervention is appropriate in designing intervention evaluations and subsequent implementations.

Machine learning methods can be used to study HTE, but different modeling techniques are needed to predict HTE than to predict overall suicide risk. HTE models can be thought of as evaluating interactions between (i) *prescriptive* predictors of intervention response (i.e., predictors of individual differences in the impact of the intervention) and (ii) either interventions or presumably causal risk

factors that could be the target of interventions in preventing subsequent outcomes. For example, if the causal effect of modifiable risk factor R was significantly stronger in leading to suicide in the presence than absence of predictor P , then P would be a prescriptive predictor of the effect of R on suicide. In this case, an intervention to prevent R would be expected to have a greater effect in reducing suicides among people with than without P .

When the effects of individual prescriptive predictors are weak and/or the number of prescriptive predictors is large, multivariate methods are needed to estimate useful HTE models. These models will sometimes have to be complex to capture nonlinear and higher order interactions of the prescriptive predictors with the target risk factor. Conventional multivariable interaction methods break down in cases of this sort and machine learning methods are required (VanderWeele et al., 2019). As discussed in more detail later in the paper, these methods can be applied even in the absence of experimental assignment to the target risk factor so long as strong predictors exist of nonrandom assignment to the target risk factor (Luedtke & van der Laan, 2017).

We illustrate this use of machine learning methods to study HTE in the current report by reanalyzing an observational dataset that documented a significant association between vitamin D deficiency and subsequent suicide in a prospective study of active-duty US military personnel (Umhau et al., 2013). Vitamin D deficiency (defined as the lowest octile of vitamin D in the sample) was found in that study to be associated with an odds ratio greater than 2.0 of subsequent suicide over a 7-year follow-up period. The rationale for attempting to study HTE in this case is as follows: Given the 12.5% prevalence of the modifiable risk factor (i.e., $1/8^{\text{th}}$ of the sample with vitamin D deficiency), an OR of 2.0 means that screening for and subsequently treating military personnel for vitamin D deficiency might prevent up to 10% of all suicides in this population. This would be cost-effective given the relatively low cost of vitamin D treatment (Singh, 2018), but screening for vitamin D deficiency can be expensive, especially when considering the possibility of screening approximately 1 million military personnel to find approximately 125,000 to treat. It might consequently make sense to focus screening efforts on those individuals found to be most at risk based on other criteria and, if HTE exists, most likely to benefit from intervention if they are found to have vitamin D deficiency. However, it would make sense to investigate the possibility of HTE before developing an intervention plan. We present a framework for doing this in the current report.

2 | METHODS

2.1 | Sample

As noted in the introduction, the example is based on a previous report documenting that vitamin D deficiency predicted subsequent suicide (Umhau et al., 2013). The sample was a 1:1 matched (on age, gender, rank, and timing of when blood samples were collected) case-

control sample of $n = 495$ suicide decedents and $n = 495$ controls from the active-duty US military with a history of combat deployment. The matched case-control design creates certain analysis complexities described below.

Detailed information on the assessment of vitamin D deficiency and the baseline measures examined as possible prescriptive predictors is presented in the original report (Umhau et al., 2013). The baseline measures included two classes of biological risk factors—continuous levels of 22 fatty acids and five trace elements—all tested in the same blood serum samples used to test for vitamin D deficiency, a series of socio-demographics and military career variables abstracted from military records, and information from medical records of psychiatric diagnoses, all as of the time of the vitamin D measurement. The categorical variables among these predictors were coded as dummies, whereas the ordinal and interval variables among the predictors were standardized (mean 0 and variance 1) and stabilized into quartiles. These transformations were important for using the full range of machine learning algorithms included in the analysis.

2.2 | Analysis methods

Overview: Numerous methods exist to estimate HTE (Robertson et al., 2020). The major challenge in the case of suicide is that even though some prescriptive predictors have been documented, none has been strong enough alone to guide precision treatment planning. This has prompted growing interest in combining information across multiple prescriptive predictors to create a composite measure of HTE (Salazar de Pablo et al., 2021). This is done most often by estimating a proportional interaction model that contains multiple interactions and combining these interactions to create a single composite measure of HTE. The latter is done using counter-factual logic: that is, by computing a predicted outcome score for each person under each intervention regimen (i.e., regardless of which regimen received) based on model coefficients and then comparing predicted individual-level regimen-specific outcome scores to select the intervention estimated to yield the better outcome for each patient (Kovalchik et al., 2013).

However, when data-driven methods are used to search for interactions, as they typically are in building composite HTE measures, there is a danger of over-fitting. Indeed, a recent simulation suggested that the majority of detected interactions in HTE models are likely to be false positives unless methods are used to reduce over-fitting (van Klaveren et al., 2019). Methods that protect against over-fitting can be used for this purpose to produce composite HTE estimates (VanderWeele et al., 2019).

With these issues in mind, it is often useful to develop an overall risk model prior to attempting to study HTE and to do so using methods that minimize risk of over-fitting by developing the model in a training sample and then testing it in an independent holdout test sample. In the case of our 495 case-control pairs, we did this by creating a training sample made up of a random 33% of observations (i.e., $n = 164$ matched case-control pairs) and a 67% testing sample

($n = 331$ pairs). We then used 10-fold cross-validation (10F-CV) in the training sample to develop several different machine learning models. Finally, we applied each model to the testing sample to evaluate the extent to which we could detect meaningful HTE. It is noteworthy that a more typical choice in a substantive context would be to use a 67% training sample and 33% test sample, as a small training sample increases risk of over-fitting.

Special issues in working with case-control samples: Case-control samples of the sort used in the vitamin D example are often used in research on suicide given the fact that suicide is a rare outcome. An analysis of a full sample, especially one based on an administrative data system, would often include thousands of non-cases for every case. Case-control analysis addresses this problem by selecting a probability sample of controls from the full set of controls either with or without weights to adjust for this under-sampling (Keogh & Cox, 2014).

The use of weights allows a wide range of models to be estimated that are appropriate for dichotomous outcomes and allows predicted risk differences to be estimated (Pedroza & Truong, 2016). When weights are not used, in comparison, the model should be estimated with logistic regression, as only the odds ratio can be estimated without bias with such a sample (Hosmer et al., 2013). The individual-level predicted odds generated by a logistic regression model can be converted into individual-level predicted risks post hoc when information is available on the probabilities of selection of controls (Rose & van der Laan, 2014). This is critical for HTE analysis, as the latter focuses on *risk differences* rather than risk ratios or odds ratios when the outcome is a dichotomy. In the example considered here, the sampling fractions used to select matched controls were not available, making it impossible to use weighting to recover predicted risks. We consequently work with logistic models and examine differences in predicted odds. We caution the reader, though, that this is being done merely to illustrate the general approach and that practical applications should use weights either prior to or subsequent to estimating the initial models and generate estimates of predicted risk rather than predicted odds.

A question might be raised on how to determine whether to weight before or after estimating a model based on a case-control design. The answer depends on the investigator's decision either to give equal weight to false positives and false negatives or, as is often the case in models for rare dichotomous outcomes, to give greater weight to detecting the rare outcome (i.e., minimizing false negatives, as when a premium is placed on detecting suicides) than to correctly classifying non-cases. Numerous methods exist for giving greater weight to detecting the rare outcome, some of them involving the use of weights but others involving either under-sampling non-cases, pseudo-sampling replicates of cases, or using a combination of both without weights (He & Ma, 2013).

Risk modeling: The first approach to estimating HTE that we investigate is known as *risk modeling*. This approach uses a 1 degree

of freedom test to evaluate the strength of HTE by generating a prediction model for the joint effects of all predictors (excluding the target risk factor) to fit a "base risk" model for the outcome in the total sample (Kent et al., 2016). The base risk is then estimated for each observation in the sample regardless of target risk factor score and used to define subgroups for investigating whether the aggregate association between the target risk factor and the outcome varies with base risk.

The intuition underlying the risk modeling approach is that some people have very low risk of the negative outcome, in which case an intervention to change a target risk factor is unlikely to have a large effect. The strength of this approach is that it provides a stable estimate of variation in aggregate outcome risk that can be evaluated with 1 degree of freedom, thereby avoiding the problem of over-fitting and often showing evidence of HTE. Consistent with this thinking, a recent secondary analysis of 32 large clinical trials (primarily in cardiology) found that most trials with significant aggregate treatment effects also had significant HTE, where the highest absolute intervention effects usually occurred among patients with the highest base risk and lowest absolute intervention effects among patients with lowest base risk (Kent et al., 2016). As a striking example, a trial of early intervention versus usual care for unstable angina found that more than half the significant aggregate treatment effect was due to an extremely strong effect among the one-eighth of patients with highest base risk and that there was no meaningful intervention effect among the 50% of patients with lowest base risk (Fox et al., 2005).

Many approaches are available to develop a base risk model, from a simple multiple regression model to a complex machine learning model. We used a stacked generalization approach based on the super learner ensemble machine learning method to develop our base risk model (Polley, 2018). In this approach, the data are analyzed in parallel in a 10F-CV training sample with a set (ensemble) of parametric and flexible prediction algorithms designed to capture nonlinearities and interactions among predictors. Results are then combined by generating a weighted composite of individual-level predicted outcome scores across the algorithms via a meta-learner equation (i.e., an equation in which the predicted outcome scores based on each algorithm is included as a separate predictor).

The advantage of this approach over others is that the composite predicted outcome score is guaranteed in expectation to perform at least as well as the best component algorithm according to a pre-specified criterion (Polley et al., 2011). We defined this as the area under the receiver operating characteristic curve (AUC), but other criteria might be preferred in other cases. Consistent with recommendations (Naimi & Balzer, 2018), we used a diverse super learner ensemble to reduce risk of misspecification (Kabir & Ludwig, 2019). These included several different linear algorithms (logistic regression, regularized regression, spline and polynomial spline regressions, support vector machines) and several different regression tree-based algorithms (boosting and bagging ensemble trees, Bayesian Additive regression trees; Table 1). All the variables in the dataset other than vitamin D status were included as potential predictors.

We applied the training sample model results to the test sample to generate individual-level predicted log-odds of suicide based on all available predictors in the dataset other than vitamin D deficiency. These predicted values were then included as the key predictor in a conditional logistic regression model that adjusted for the matching of cases and controls (Sun et al., 2011) and included a dummy variable for vitamin D deficiency, the main effect of the predicted values, and an interaction between vitamin D deficiency and the predicted values. The existence of HTE was determined by evaluating the significance of the interaction. We also examined the possibility of nonlinearities by estimating a separate conditional logistic regression model in which dummy variables were created for quartiles of the predicted log-odds and interactions were estimated between the dummy for vitamin D deficiency and the log-odds dummies.

Lasso penalized logistic regression: As noted above, the conventional approach to HTE estimation is the proportional interaction model (Kovalchik et al., 2013). Such a model can be estimated by including all potential prescriptive predictors, the target risk factor, and two-way interactions between the target risk factor and the potential prescriptive predictors. The existence of HTE is evaluated by testing the significance of the interactions. However, as such a model will almost certainly overfit the data (van Klaveren et al., 2019), it is usually wise to use some type of penalty to minimize risk of over-fitting. One way to do this is with Least Absolute Shrinkage and Selection Operator (Lasso) penalized regression. Lasso performs both variable selection and regularization by forcing the sum of the absolute standardized values of all regression coefficients in a model to be less than

some fixed value associated to the regularization penalty, thereby forcing some coefficients to zero and using internal CV to determine the optimal penalty value (Tibshirani, 1996). This reduces risk of over-fitting. This is the second approach we investigated in our illustrative analysis. Specifically, we used lasso to select a small set of stable interactions based on 10F-CV in the training sample of our dataset to estimate a reduced conditional logistic model in the test sample. We then searched for HTE by evaluating the significance of the coefficients in this proportional interaction specification.

Causal forest: Although CV can be used to minimize the problem of over-fitting (Abadie et al., 2018), the lasso penalized regression approach, like many other approaches for estimating HTE, can be faulted because accuracy still requires correct specification of both the (possibly nonlinear) main effects and the (possibly complex nonlinear and higher order) interactions. However, other algorithms exist that estimate interactions directly and do not require correct specification of the main effects although they do require correct specification of the interaction terms (Pan & Zhao, 2021; Wang et al., 2018). The third approach we investigate in our illustrative analysis uses one of the most recently developed of these algorithms of this sort: *causal forest* (Wager & Athey, 2018). This algorithm is an extension of *random forests*, an algorithm that averages predicted outcome values over many classification trees, each based on a subsample of predictors, to correct for the problem of over-fitting in more conventional regression trees (Breiman, 2001). The causal forest algorithm uses the same logic, but rather than splitting to minimize prediction error in an outcome, it splits to maximize differences in the

TABLE 1 Algorithms used in the super learner ensemble^a

Algorithm	Description
I. Super learner	Super learner is an ensemble machine learning approach that uses cross-validation (CV) to select a weighted combination of predicted outcome scores across a collection of candidate algorithms (learners) to yield an optimal combination according to a pre-specified criterion that performs at least as well as the best component algorithm. R package: <i>Super learner</i> (van der Laan et al., 2007).
II. Linear algorithms in the super learner library	
A. Generalized linear models	Maximum likelihood estimation with logistic link function. R package: <i>stats</i> (Nelder & Wedderburn, 1972).
B. Elastic Net	Elastic net is a regularization method that minimizes the problem of overlap among predictors by explicitly penalizing over-fitting with a composite penalty $\lambda \{MPP \times \text{Plasso} + (1 - MPP) \times \text{Pridge}\}$, where MPP is a mixing parameter penalty with values between 0 and 1 that controls relative weighting between the lasso penalty (Plasso) and the ridge penalty (Pridge). The parameter λ controls the total amount of penalization. The ridge penalty handles multicollinearity by shrinking all coefficients smoothly towards 0 but retains all variables in the model. The lasso penalty allows simultaneous coefficient shrinkage and variable selection, tending to select at most one predictor in each strongly correlated set, but at the expense of giving unstable estimates in the presence of high multicollinearity. The elastic net approach of combining the ridge and lasso penalties has the advantage of yielding more stable and accurate estimates than either ridge or lasso alone while maintaining model parsimony. R package: <i>glmnet</i> (Friedman et al., 2010). 11 different <i>glmnet</i> specifications were used,

(Continues)

TABLE 1 (Continued)

Algorithm	Description
C. Adaptive splines	<p>varying the α hyperparameter over the values 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, and 1.</p> <p>Adaptive spline regression flexibly captures both linear and piecewise nonlinear associations as well as interactions among these associations by connecting linear segments (splines) of varying slopes and smooths to create piece-wise curves (basis functions). Final fit is built using a stepwise procedure that selects the optimal combination of basic functions. <i>R</i> package: <i>Earth</i> (Milborrow, 2021). 3 different <i>Earth</i> specifications were used, varying the degree hyperparameter over the values 1, 3, and 5.</p>
D. Adaptive polynomial splines ^b	<p>Adaptive polynomial splines are like adaptive splines but differ in the order in which basis functions (e.g., linear vs. nonlinear) are added to build the final model. <i>R</i> package: <i>polspline</i>; Kooperberg, 2020.</p>
III. Tree-based algorithms	
A. Bagging	<p>Random forest. Independent variables are partitioned (based on contiguous values) and stacked to build short decision trees that are combined (ensemble) to create an aggregate "forest". Random forest builds numerous trees in bootstrapped samples and generates an aggregate tree by averaging across trees, thereby reducing over-fitting. <i>R</i> package: <i>ranger</i> (Wright & Ziegler, 2017). 3 different <i>ranger</i> specifications were used, each with the following hyperparameter values: <i>max.depth</i> = (6, 8, 8), <i>num.trees</i> = (1500, 1700, 1000), <i>mtry</i> = (10, 4, 20), <i>splitrule</i> = ("gini," "hellinger," "extratrees")</p>
B. Gradient boosting	<p>Gradient boosting algorithms build a sequential ensemble of shallow successive regression trees that iteratively learn the residuals from prior trees. This is a flexible method, where the number of trees, interaction depth, and shrinkage are leveraged to build flexible models. <i>R</i> package: <i>CatBoost</i> (Prokhorenkova et al., 2019). 2 different <i>CatBoost</i> specifications were used, each with the following hyperparameter values: <i>Iterations</i> = (50, 100), <i>learning_rate</i> = (0.3, 0.8), <i>depth</i> = (8, 10).</p>
C. Extreme gradient boosting	<p>A fast and efficient implementation of gradient boosting. <i>R</i> package: <i>XGBoost</i> (Chen & Guestrin, 2016). 5 different <i>XGBoost</i> specifications were used, each with the following hyperparameters: <i>Ntrees</i> = (1000, 100, 500, 100, 800), <i>max_depth</i> = (6, 2, 6, 8, 4), <i>shrinkage</i> = (0.001, 0.1, 0.1, 0.1, 0.001), <i>gamma</i> = (0.3, 0.5, 0.0, 0.5, 0.8), <i>minobspnode</i> = (20, 10, 20, 10, 20), and <i>colsample_bytree</i> = (0.3, 0.8, 0.5, 0.3, 0.8)</p>
D. DBARTS	<p>Fits Bayesian additive regression trees. <i>R</i> package: <i>dbarts</i>. (Dorie, 2020).</p>

Abbreviation: DBARTS, Discrete Bayesian Additive Regression Trees Sampler.

^aEach linear algorithm was estimated separately with five different lasso screeners where *dfmax* = 10, 15, 20, 30 and all predictors. Each tree algorithm was estimated separately with five different *ranger* screeners for number of predictors equal to 10, 15, 20 30 and all predictors. Hyperparameter tuning was achieved by treating different specifications of individual algorithms as separate learners in the ensemble, as detailed in the body of the table.

^bHyperparameters: Default values were used unless otherwise noted.

association between a target dichotomous risk factor and an outcome across subgroups. The output is a predicted slope defined as the individual's odds-ratio of the outcome in the presence versus absence of the target risk factor. Importantly, this estimate is independent of whether the individual does or does not have the target risk factor, as the estimate is based on counter-factual logic in which each person is implicitly assigned two estimated outcomes—one in the presence and the other in the absence of the target risk factor. The individual's estimated HTE is the difference between these two predicted values.

In addition to applying the causal forest algorithm to our training sample, we calculated SHapley Additive exPlanations (SHAP) values to estimate the relative importance of each predictor variable in the model in the 10F-CV training sample (Lundberg & Lee, 2017). SHAP values represent the marginal contribution to overall model accuracy

of each variable in a predictor set. The causal forest model was then re-estimated with only the top five and then the top 25 predictors defined by SHAP values to evaluate the effects of over-fitting on model results. The 3 causal forest models estimated in the training sample were then used to generate 3 separate sets of individual-level predicted odds-ratios in the test sample. The existence of HTE under each of these specifications was determined by evaluating the significance of the interaction between the dummy variable for vitamin D deficiency and the predicted individual-level odds-ratios for reactivity to this deficiency.

Data management and estimation of the conditional logistic models were carried out in SAS version 9.4 (SAS Institute Inc., 2013). The lasso, super learner, and causal forest models were estimated in *R* version 3.6.3 (R Core Team, 2020). SHAP values were estimated in Python (Lundberg, 2018).

3 | RESULTS

3.1 | The aggregate association

The aggregate association (95% CI) of vitamin D deficiency with suicide in the total case-control sample was $OR = 2.1$ (1.3–3.2). This association was statistically significant ($\chi^2_1 = 9.9$, $p < 0.002$). The association was weaker, though, in the randomly selected 67% test sample, $OR = 1.4$ (0.7–2.6), $\chi^2_1 = 1.1$, $p = 0.30$. It is noteworthy, though, that it is useful in practical applications to design training and test sample splits so that the distribution of the target risk factor, the outcome, and the association between the two is the same.

3.2 | Using super learner to generate a risk model

A total of 334 variables other than vitamin D deficiency were included in the overall dataset. All these variables were included as potential predictors in the super learner ensemble in the training sample. The algorithms with the highest super learner weights were the adaptive splines, with the linear splines accounting for 35.4% of total ensemble weight and the polynomial splines for an additional 48.9% (Table 2). The other additive algorithms (9.0%) and tree-based algorithms (6.6%) were much less important. The AUC (standard error) of the full super learner ensemble model when applied to the independent test sample was $AUC = 0.76$ (0.02).

The interaction of the dummy variable for vitamin D deficiency with the standardized (to a mean of 0 and variance of 1) continuous overall super learner predicted odds of suicide in the test sample was negative and nonsignificant: $OR = 0.2$ (95% CI:0.0–4.0, $\chi^2_1 = 1.0$, $p = 0.32$; Table 3). The OR being less than 1.0 means that the increased relative-odds of suicide associated with vitamin D deficiency decreases as overall super learner predicted odds increases. This pattern can be seen clearly in Model 2, where we divided the continuous predicted odds into quartiles. The ORs associated with the 3 dummy variables for the upper three quartiles (Q2–Q4) compared to those in the lowest quartile (Q1) can be interpreted as conditional ORs among people who do not have vitamin D deficiency.

Consistent with previous machine learning models predicting suicide (Burke et al., 2019), the OR is dramatically higher in the highest quartile $OR = 62.3$ (18.5–209.8, $\chi^2_1 = 44.5$, $p < 0.001$) than in intermediate quartiles ($OR = 1.2$ –1.9) or the lowest quartile (where the OR is implicitly 1.0). The OR for vitamin D deficiency in Model 2, $OR = 2.7$ (0.8–9.1, $\chi^2_1 = 2.6$, $p = 0.11$), can be interpreted as the OR among individuals in the lowest quartile of super learner predicted odds. This OR is somewhat higher than in the total sample: $OR = 2.1$. The interactions in Model 2 are 0.5 for the two intermediate quartiles and 0.2 for the quartile with highest super learner predicted odds. Which means that the ORs are approximately $OR = 1.3$ (i.e., 0.5×2.7) in the half-sample with intermediate super learner predicted odds and $OR = 0.5$ (i.e., 0.2×2.7) in the quartile with highest super learner predicted odds. Although nonsignificant ($\chi^2_3 = 1.8$,

$p = 0.60$), these negative interactions are nonetheless noteworthy for reasons described in the discussion section.

3.3 | The lasso penalized logistic regression model

The lasso model was estimated with main effects for vitamin D deficiency and each of the 334 other variables included in the super learner model along with a separate interaction of vitamin D deficiency with each of these 334 variables. A total of 23 main effects and four interactions were retained in the final lasso model (Table A1). We estimated four different conditional logistic models in the test sample to evaluate the significance of these interactions as evidence of HTE (Table 4). The first model included only the dummy variable for vitamin D deficiency, main effects of the four variables with interactions in the lasso model, and the interactions of vitamin D deficiency with these four variables. The second model added the 23 other variables that had main effects in the lasso model. The third and fourth models deleted the main effects of the four interacting variables, none of which was in the lasso model. The fourth model also deleted the main effect of vitamin D deficiency, which was not part of the lasso model. None of the interactions was significant either singly or as a set in any of these models.

Had one or more of these models been significant, though, the next step in the analysis would have been to carry out a simulation in which we estimated two predicted outcome scores for each individual in the sample: based on the assumption that the target risk factor was either positive (i.e., the individual had low vitamin D) or negative. Individual-level comparisons would then be made to determine which individuals had the highest difference scores. This would have been done using 10F-CV to minimize over-estimation of the difference scores. Information external to the model regarding the costs, effectiveness, and competing risks of treatment would then have been combined with this information about predicted difference scores to determine the decision threshold for intervention. Decision science methods exist for making principled decisions of this sort when individual-level estimates of intervention effects exists (Kinchin et al., 2017; Van Calster et al., 2018).

3.4 | The causal forest algorithm

The same 334 variables were used as predictors in the causal forest analysis in the training sample. One hundred forty-nine of these variables had nonzero SHAP variable importance values in the training sample. However, as it is typically the case in such analyses, most of these variables had small SHAP values (135 less than 0.001). As a result, in an effort to evaluate the possibility of overfitting, we repeated the causal forest analysis with only the five and 25 predictors with highest SHAP values. Consistent with the fact that SHAP values decreased markedly after the few most important predictors, Pearson correlations were extremely high between the causal forest predicted difference scores based on all variables and based on the

TABLE 2 Nonzero super learner algorithm weights in the training sample

	Feature selection	Hyperparameter Tuning	Weight
I. Linear algorithms			
Generalized linear model	All		0.9%
Elastic net	15	$a = 0$	2.6%
	20	$a = 0$	5.5%
Adaptive splines	All	degree = 1	20.6%
	All	degree = 3	9.1%
	All	degree = 5	5.8%
Adaptive polynomial splines	10		33.9%
	15		15.0%
Total linear	-	-	93.4%
II. Tree-based algorithms			
Extreme gradient boosting	10	#3 ^a	1.3%
	10	#5 ^a	0.1%
DBARTS	10		4.8%
	All		0.5%
Total tree-based	-	-	6.7%

Abbreviations: DBARTS, Discrete Bayesian Additive Regression Trees Sampler

^aThe 3rd and 5th specifications in Table 1.

	Model 1			Model 2		
	OR	(95% CI)	χ^2_1	OR	(95% CI)	χ^2_1
Main effects						
Vitamin D deficiency	1.4	(0.7–2.6)	1.1	2.7	(0.8–9.0)	2.6
SL predicted odds ^b						
Continuous	3.9 ^a	(2.8–5.3)	72.3	-	-	-
Q1				1.0		
Q2	-	-	-	1.1	(0.6–2.0)	0.2
Q3	-	-	-	1.9 ^a	(1.0–3.5)	4.3
Q4 (highest)	-	-	-	62.2 ^a	(18.5–209.6)	44.5
χ^2_3	-	-		46.8 ^a		
Interactions						
SL predicted odds ^b						
Continuous	0.7	(0.3–1.4)	1.0	-	-	-
Q1				1.0		
Q2	-	-	-	0.5	(0.1–2.5)	0.6
Q3	-	-	-	0.4	(0.1–2.0)	1.1
Q4 (highest)	-	-	-	0.2	(0.0–2.8)	1.4
χ^2_3	-	-		1.8		

Abbreviations: CI, confidence interval; OR, odds ratio; SL, super learner.

^aSignificant at the 0.05 level, two-sided test.

^bPredicted odds from the SL model was standardized in the test sample.

TABLE 3 Interactions between vitamin D deficiency and the super learner estimate of composite predicted odds (developed in the training sample) in predicting subsequent suicide based on a conditional logistic regression model estimated in the test sample ($n = 331$ matched pairs)

TABLE 4 Interactions between vitamin D deficiency and four variables selected by Least Absolute Shrinkage and Selection Operator (Lasso; in the training sample) in predicting subsequent suicide based on a conditional logistic regression model estimated in the test sample ($n = 662$; matched pairs = 331)^a

	Model 1		Model 2		Model 3		Model 4	
	OR	(95% CI)	OR	(95% CI)	OR	(95% CI)	OR	(95% CI)
FA cluster: Risky versus protective ^b	1.0	(0.3–3.1)	0.6	(0.1–2.4)	0.7	(0.2–2.5)	0.9	(0.4–1.9)
Rank: Officer ^c	-	-	-	-	-	-	-	-
Percent of DGLA ^d	1.0	(0.6–1.7)	1.0	(0.5–1.9)	0.8	(0.4–1.6)	0.8	(0.4–1.6)
Ratio of stearic acid to palmitic acid ^d	0.7	(0.4–1.2)	0.5*	(0.2–1.0)	0.5*	(0.3–1.0)	0.5*	(0.3–1.0)
χ^2_3	2.3 ^e		4.4 ^e		4.8 ^e		4.8 ^e	

Abbreviations: CI, confidence interval; DGLA, dihomo- γ -linolenic acid; FA, fatty acid; OR, odds ratio.

^aModel 1 included only the dummy variable for vitamin D deficiency, main effects of the 4 variables with interactions in the LASSO model, and the interactions of vitamin D deficiency with these 4 variables as predictors. Only interaction coefficients are shown here. Model 2 added controls for the 23 other variables with main effects in the LASSO model. Models 3 and 4 deleted the main effects of the four interacting variables, none of which was in the LASSO model, whereas Model 4 additionally deleted the main effect of vitamin D deficiency, which was not in the LASSO model.

^bRisky and protective fatty acid clusters were defined based on the clusters discovered by Ryan et al., 2021.

^cOnly $n = 8$ of the $n = 52$ officers in the test sample had vitamin D deficiency. Seven of these eight were suicide cases. This compares to $n = 19$ cases and $n = 25$ controls among officers without vitamin D deficiency, for a gross OR of 9.2. The comparable gross OR among others in the sample (i.e., those that were not officers) was 1.1 ($n = 46$ cases and $n = 41$ controls among those with vitamin D deficiency; $n = 259$ cases and $n = 264$ controls among those without vitamin D deficiency), resulting in a gross interaction OR of 8.3. However, this coefficient became unstable in the multivariate model and could not be estimated. It is noteworthy that the comparable OR in the training sample LASSO model had the opposite sign (OR = 0.9).

^dStandardized variable.

^eNo χ^2 tests were significant ($p = 0.31$ – 0.68).

*Significant at the 0.05 level, two-sided test.

top 25 ($r = 0.98$) and top five variables ($r = 0.92$) as well as between scores based on the top 25 and top five variables ($r = 0.91$).

We estimated a separate conditional logistic model in the test sample for each of these three scores, in each case including the main effect of vitamin D deficiency, the main effect of the causal forest score, and the interaction between the two variables. The interaction was nonsignificant in all three cases ($\chi^2_1 = 0.5$ – 1.0 , $p = 0.49$ – 0.33). As with the super learner risk model, we then divided each causal forest score into quartiles to inspect conditional ORs of vitamin D deficiency with subsequent suicide (Table 5). Variation in ORs across quartiles was nonmonotonic. Had the interaction been significant, we would have applied the same principled methods for evaluating a clinical decision threshold as described above in the discussion of the lasso model.

4 | DISCUSSION

We focused above on the logic of HTE analysis and application of some commonly used algorithms for this type of analysis. However, it is important to note that we implicitly assumed in all these modeling efforts that the target predictor variable was causal and was randomly assigned with respect to the potential prescriptive predictors. Whereas these assumptions are plausible when applied to an experiment in which we manipulate exposure to the presumed causal risk factor, it is not a plausible assumption for most observational studies. Failure to take this into account can lead to biased estimates of HTE (Mozer et al., 2020). However, it often occurs, as in the

example considered here, that we are interested in evaluating the possible existence of HTE as a preliminary to carrying out an experiment. How might that be done?

As it happens, principled methods exist to work with observational data to make preliminary estimates of HTE if the baseline predictor set includes the important determinants of the subset of predictors of nonrandom exposure to the target risk factor that are also independent causes of the outcome, as the differences in these baseline covariates can be “balanced” statistically to approximate the distributions found in experimental trials (Hirshberg & Zubizarreta, 2017; Visconti & Zubizarreta, 2018). It has been shown that analyses of such balanced databases often yield aggregate results very similar to those obtained in experimental trials (Anglemyer et al., 2014; Dahabreh et al., 2012).

It will often happen, though, that some confounders are unmeasured. When this is the case, it is sometimes possible to find natural variation that mimics an experiment (Handley et al., 2018). For example, opportunities of this sort could exist to study policy interventions of various sorts using before-after ecological designs, such as the aggregate effects of interventions to reduce suicide by means of restrictions of various kinds (Zalsman et al., 2016). Analysis could make use of a regression discontinuity design (Moscoe et al., 2015; Venkataramani et al., 2016). Or it might be possible to make principled causal inferences about aggregate treatment effects by attempting to find an instrumental variable (IV; Baiocchi et al., 2014; Swanson, 2017), where a known cause of the target risk factor is measured that we are willing to assume affects the outcome only through the intervention.

	Model 1 (All)		Model 2 (Top 25)		Model 3 (Top 5)	
	OR	(95% CI)	OR	(95% CI)	OR	(95% CI)
Q1 (lowest) protective	1.1	(0.5–2.6)	1.0	(0.4–2.5)	1.0	(0.4–2.4)
Q2	1.5	(0.6–3.8)	1.6	(0.7–3.9)	1.3	(0.5–3.4)
Q3	1.1	(0.4–2.9)	1.0	(0.3–3.1)	4.0*	(1.3–12.7)
Q4 (highest) palmitic acid	1.8	(0.7–4.8)	1.5	(0.6–3.7)	0.8	(0.3–2.0)
χ^2_3	2.4 ^b		2.2 ^b		6.2 ^b	

Abbreviations: CI, confidence interval; OR, odds ratio.

^aModel 1 was based on the causal forest algorithm that used all 149 predictors with nonzero variable importance values in the training sample. Model 2 was based on a separate causal forest model that used only the 25 predictors with highest SHAP values in Model 1. Model 3 was based on a separate causal forest model that used only the five predictors with highest SHAP values in Model 1.

^bNo χ^2 tests were significant ($p = 0.18$ – 0.70).

*Significant at the 0.05 level, two-sided test.

We know, for example, that vitamin D deficiency varies with latitude and with season (Leary et al., 2017) and it might be reasonable in some circumstances to assume that these are valid instrumental variables. When this is the case, both observed and unobserved covariates across groups defined by the instrumental variable can be balanced to identify the effect of a potential intervention. Both regression discontinuity and IV analyses can be conducted with other covariate balancing methods (Keele et al., 2015; Zubizarreta et al., 2013) and combined with additional regression adjustments using more complex estimators (Robins & Rotnitzky, 1995).

Second, whereas we investigated only a small series of HTE modeling approaches, a great many different algorithms exist to estimate HTE and none of them is optimal for all applications. The reason is that different ways of estimating interactions differ in their assumptions, and the best method for any given application will be the one whose assumptions conform best to the true structure of the interactions in the population for that application. Some researchers attempt to address this problem by carrying out analyses using several different approaches, as we did here, and then selecting the approach that has the best CV results as the one they use in application. However, a better way to proceed is to use a stacked generalization approach in which results are combined (rather than compared) across multiple algorithms. This is the approach we recommend using.

Another issue concerns the choice among the many methods for estimating HTE. We described several such algorithms. Many others exist. Rather than try to decide on a preferred algorithm or compare a handful and select the best one out of those compared, we prefer to use stacked generalization to generate a single model to be estimated that combines results across many different algorithms. As described above in the discussion of super learner, this is done by generating a weight that combines predictions optimally across the algorithms. Stacked generalization makes the final model less prone to misspecification than approaches based on a single algorithm (van der

TABLE 5 Within-quartile associations between vitamin D deficiency and subsequent suicide based on quartiles of the causal forest estimate of log-odds differences (developed in the training sample) based on a conditional logistic regression model estimated in the test sample ($n = 662$; matched pairs = 331)^a

Laan & Luedtke, 2015). Stacked generalization can be carried out in conjunction with weighting or matching to adjust for measured confounders (Luedtke & van der Laan, 2017) and with regression discontinuity or IV designs to adjust for unmeasured confounders (Qiu et al., 2021).

Finally, in the case where HTE is being estimated from observational data, it is important to emphasize the role of separating the design and analysis stages. Following Rubin (2008), all the data adjustments, empirical evaluations, and fitted models that do not require outcome information correspond to the design stage of an observational study, while all the examinations that use this information belong to the analysis stage. Separating these two stages is important because it helps preserve the study's objectivity and maintain the validity of its statistical tests. Our study followed this principle by first learning the relevant variables for HTE in a training sample, and then fitting the final effect models in another disjoint part of the total sample. On the learning part of the data, we leveraged a variety of machine learning approaches to inform the final causal effect models on the analysis part. At a high level, this also allowed us to integrate modern ideas from machine learning with classical procedures for observational studies, such as conditional logistic regression models for case-control studies. These latter models also catalyzed our substantive knowledge of the problem under study. Moving forward, the general approach of using flexible optimization-based algorithms for prediction guided by study design principles for causal inference that integrate substantive knowledge of the problem at hand is a promising route for future HTE analysis.

ACKNOWLEDGMENTS

The authors thank Dr. J. R. Hibbeln for his contributions to the initiation of the original project and specifically the creation of the original dataset analyzed in this study, and current comments on the manuscript; A. Dagdag and T. Stubborn for their steadfast support in the final progression of this project; and Iqra Mohyuddin for her help with the data management of the final dataset. Dr. Jose R.

Zubizarreta was supported through a Patient-Centered Outcomes Research Institute (PCORI) Project Program Award (ME-2019C1-16172). The initial data collection, management of statistical analysis and interpretation of data were supported by the Division of Intramural Basic and Clinical Research, NIAAA/ NIH and the Defense Advanced project Agency, Arlington, Virginia. Additional support for Drs. Teodor T. Postolache and Lisa A. Brenner's participation in this project was received from the Rocky Mountain MIRECC for Suicide Prevention, Aurora, Colorado.

CONFLICT OF INTEREST

In the past 3 years, Dr. Ronald C. Kessler was a consultant for Datastat, Inc., Holmusk, RallyPoint Networks, Inc., and Sage Therapeutics. He has stock options in Mirah, PYM, and Roga Sciences. The other authors declare no conflicts of interest.

DATA AVAILABILITY STATEMENT

Data cannot be shared directly given their nature and the previous administrative approvals of the project. Those with questions about potential additional analyses of the dataset or collaborations should contact Dr. Patricia A. Deuster at the Department of Military and Emergency Medicine, Uniformed Services University of Health Sciences at patricia.deuster@usuhs.edu.

DISCLAIMER

The opinions and assertions expressed herein are those of the authors and do not reflect the official policy or position of the Uniformed Services University, National Institutes of Health, the Department of Defense or the Veterans Health Administration.

ORCID

Ronald C. Kessler  <https://orcid.org/0000-0003-4831-2305>

REFERENCES

- Abadie, A., Chingos, M. M., & West, M. R. (2018). Endogenous stratification in randomized experiments. *The Review of Economics and Statistics*, 100(4), 567–580. https://doi.org/10.1162/rest_a_00732
- Anglemyer, A., Horvath, H. T., & Bero, L. (2014). Healthcare outcomes assessed with observational study designs compared with those assessed in randomized trials. *Cochrane Database of Systematic Reviews*, 2014(4). <https://doi.org/10.1002/14651858.MR000034.pub2>
- Baiocchi, M., Cheng, J., & Small, D. S. (2014). Instrumental variable methods for causal inference. *Statistics in Medicine*, 33(13), 2297–2340. <https://doi.org/10.1002/sim.6128>
- Barak-Corren, Y., Castro, V. M., Javitt, S., Hoffnagle, A. G., Dai, Y., Perlis, R. H., Nock, M. K., Smoller, J. W., & Reis, B. Y. (2017). Predicting suicidal behavior from longitudinal electronic health records. *American Journal of Psychiatry*, 174(2), 154–162. <https://doi.org/10.1176/appi.ajp.2016.16010077>
- Belsher, B. E., Smolenski, D. J., Pruitt, L. D., Bush, N. E., Beech, E. H., Workman, D. E., Morgan, R. L., Evatt, D. P., Tucker, J., & Skopp, N. A. (2019). Prediction models for suicide attempts and deaths: A systematic review and simulation. *JAMA Psychiatry*, 76(6), 642–651. <https://doi.org/10.1001/jamapsychiatry.2019.0174>
- Bossarte, R. M., Kennedy, C. J., Luedtke, A., Nock, M. K., Smoller, J. W., Stokes, C., & Kessler, R. C. (2021). New directions in machine learning analyses of administrative data to prevent suicide-related behaviors. *American Journal of Epidemiology*. Advance online publication. <https://doi.org/10.1093/aje/kwab111>
- Breiman, L. (2001). Random forests. *Machine Learning*, 45(1), 5–32. <https://doi.org/10.1023/A:1010933404324>
- Brodsky, B. S., Spruch-Feiner, A., & Stanley, B. (2018). The zero suicide model: Applying evidence-based suicide prevention practices to clinical care. *Frontiers in Psychiatry*, 9, 33. <https://doi.org/10.3389/fpsy.2018.00033>
- Burke, T. A., Ammerman, B. A., & Jacobucci, R. (2019). The use of machine learning in the study of suicidal and non-suicidal self-injurious thoughts and behaviors: A systematic review. *Journal of Affective Disorders*, 245, 869–884. <https://doi.org/10.1016/j.jad.2018.11.073>
- Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 785–794). Association for Computing Machinery. <https://doi.org/10.1145/2939672.2939785>
- Comtois, K. A., Kerbrat, A. H., DeCou, C. R., Atkins, D. C., Majeres, J. J., Baker, J. C., & Ries, R. K. (2019). Effect of augmenting standard care for military personnel with brief caring text messages for suicide prevention: A randomized clinical trial. *JAMA Psychiatry*, 76(5), 474–483. <https://doi.org/10.1001/jamapsychiatry.2018.4530>
- Dahabreh, I. J., Sheldrick, R. C., Paulus, J. K., Chung, M., Varvarigou, V., Jafri, H., Rassen, J. A., Trikalinos, T. A., & Kitsios, G. D. (2012). Do observational studies using propensity score methods agree with randomized trials? A systematic comparison of studies on acute coronary syndromes. *European Heart Journal*, 33(15), 1893–1901. <https://doi.org/10.1093/eurheartj/ehs114>
- Dorie, V. (2020). *Discrete Bayesian additive regression trees sampler*. Retrieved from <https://cran.r-project.org/web/packages/dbarts/dbarts.pdf>
- Fernandes, B. S., Williams, L. M., Steiner, J., Leboyer, M., Carvalho, A. F., & Berk, M. (2017). The new field of 'precision psychiatry'. *BMC Medicine*, 15(1), 80. <https://doi.org/10.1186/s12916-017-0849-x>
- Fox, K. A., Poole-Wilson, P., Clayton, T. C., Henderson, R. A., Shaw, T. R., Wheatley, D. J., Knight, R., & Pocock, S. J. (2005). 5-year outcome of an interventional strategy in non-ST-elevation acute coronary syndrome: The British Heart Foundation RITA 3 randomised trial. *Lancet*, 366(9489), 914–920. [https://doi.org/10.1016/s0140-6736\(05\)67222-4](https://doi.org/10.1016/s0140-6736(05)67222-4)
- Friedman, J. H., Hastie, T., & Tibshirani, R. (2010). Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software*, 33(1), 22. <https://doi.org/10.18637/jss.v033.i01>
- Gradus, J. L., Rosellini, A. J., Horváth-Puhó, E., Street, A. E., Galatzer-Levy, I., Jiang, T., Lash, T. L., & Sørensen, H. T. (2020). Prediction of sex-specific suicide risk using Machine Learning and Single-Payer health care registry data from Denmark. *JAMA Psychiatry*, 77(1), 25–34. <https://doi.org/10.1001/jamapsychiatry.2019.2905>
- Handley, M. A., Lyles, C. R., McCulloch, C., & Cattamanchi, A. (2018). Selecting and improving quasi-experimental designs in effectiveness and implementation research. *Annual Review of Public Health*, 39, 5–25. <https://doi.org/10.1146/annurev-publhealth-040617-014128>
- He, H., & Ma, Y. (2013). *Imbalanced learning: Foundations, algorithms, and applications*. Wiley-IEEE Press. <https://doi.org/10.1002/9781118546106>
- Hirshberg, D. A., & Zubizarreta, J. R. (2017). On two approaches to weighting in causal inference. *Epidemiology*, 28(6), 812–816. <https://doi.org/10.1097/ede.0000000000000735>
- Hoge, C. W. (2019). Suicide reduction and research efforts in Service Members and Veterans-Sobering realities. *JAMA Psychiatry*, 76(5), 464–466. <https://doi.org/10.1001/jamapsychiatry.2018.4564>
- Hosmer, D. W., Jr, Lemeshow, S., & Sturdivant, R. X. (2013). *Applied logistic regression* (3rd ed.). John Wiley & Sons. <https://doi.org/10.1002/9781118548387>
- Jobes, D. A. (2012). The Collaborative Assessment and Management of Suicidality (CAMS): An evolving evidence-based clinical approach to

- suicidal risk. *Suicide and Life-Threatening Behavior*, 42(6), 640–653. <https://doi.org/10.1111/j.1943-278X.2012.00119.x>
- Kabir, M. F., & Ludwig, S. A. (2019). Enhancing the performance of classification using super learning. *Data-Enabled Discovery and Applications*, 3(1), 5. <https://doi.org/10.1007/s41688-019-0030-0>
- Keele, L., Titiunik, R., & Zubizarreta, J. R. (2015). Enhancing a geographic regression discontinuity design through matching to estimate the effect of ballot initiatives on voter turnout. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 178(1), 223–239. <https://doi.org/10.1111/rssa.12056>
- Kent, D. M., Nelson, J., Dahabreh, I. J., Rothwell, P. M., Altman, D. G., & Hayward, R. A. (2016). Risk and treatment effect heterogeneity: Re-analysis of individual participant data from 32 large clinical trials. *International Journal of Epidemiology*, 45(6), 2075–2088. <https://doi.org/10.1093/ije/dyw118>
- Keogh, R. H., & Cox, D. R. (2014). *Case-control studies*. Cambridge University Press. <https://doi.org/10.1017/CBO9781139094757>
- Kessler, R. C., Bernecker, S. L., Bossarte, R. M., Luedtke, A. R., McCarthy, J. F., Nock, M. K., Pigeon, W. R., Petukhova, M. V., Sadikova, E., VanderWeele, T. J., Zuromski, K. L., & Zaslavsky, A. M. (2019). The role of big data analytics in predicting suicide. In I. C. Passos, B. Mwangi, & F. Kapczinski (Eds.), *Personalized psychiatry* (pp. 77–98). Springer.
- Kessler, R. C., Bossarte, R. M., Luedtke, A., Zaslavsky, A. M., & Zubizarreta, J. R. (2020). Suicide prediction models: A critical review of recent research with recommendations for the way forward. *Molecular Psychiatry*, 25(1), 168–179. <https://doi.org/10.1038/s41380-019-0531-0>
- Kinchin, I., Doran, C. M., Hall, W. D., & Meurk, C. (2017). Understanding the true economic impact of self-harming behaviour. *Lancet Psychiatry*, 4(12), 900–901. [https://doi.org/10.1016/s2215-0366\(17\)30411-x](https://doi.org/10.1016/s2215-0366(17)30411-x)
- Kooperberg, C. (2020). *Polynomial spline routines*. Retrieved from <https://cran.r-project.org/web/packages/polspline/polspline.pdf>
- Kovalchik, S. A., Varadhan, R., & Weiss, C. O. (2013). Assessing heterogeneity of treatment effect in a clinical trial with the proportional interactions model. *Statistics in Medicine*, 32(28), 4906–4923. <https://doi.org/10.1002/sim.5881>
- Leary, P. F., Zamfirova, I., Au, J., & McCracken, W. H. (2017). Effect of latitude on vitamin D levels. *Journal of the American Osteopathic Association*, 117(7), 433–439. <https://doi.org/10.7556/jaoa.2017.089>
- Luedtke, A. R., & van der Laan, M. J. (2017). Evaluating the impact of treating the optimal subgroup. *Statistical Methods in Medical Research*, 26(4), 1630–1640. <https://doi.org/10.1177/0962280217708664>
- Lundberg, S. (2018). *Welcome to the SHAP documentation*. Retrieved from <https://shap.readthedocs.io/en/latest/index.html>
- Lundberg, S. M., & Lee, S.-I. (2017). Advances in Neural Information Processing Systems 30: 31st Annual Conference on Neural Information Processing Systems (NIPS 2017). In I. Guyon, U. von Luxburg, S. Bengio, H. M. Wallach, R. Fergus, S. V. N. Vishwanathan, & R. Garnett (Eds.), *Advances in Neural Information Processing Systems* (Vol. 30, pp. 4768–4777). NeurIPS Proceedings.
- Milborrow, S. (2021). *Multivariate adaptive regression splines*. Retrieved from <https://cran.r-project.org/web/packages/earth/earth.pdf>
- Moscoe, E., Bor, J., & Bärnighausen, T. (2015). Regression discontinuity designs are underutilized in medicine, epidemiology, and public health: A review of current and best practice. *Journal of Clinical Epidemiology*, 68(2), 132–143. <https://doi.org/10.1016/j.jclinepi.2014.06.021>
- Mozer, R., Rubin, D. B., & Zubizarreta, J. (2020). Statistical inference for causal effects in clinical psychology: Fundamental concepts and analytical approaches. In A. G. C. Wright & M. N. Hallquist (Eds.), *The Cambridge Handbook of Research Methods in Clinical Psychology* (pp. 415–425). Cambridge University Press. <https://doi.org/10.1017/9781316995808.038>
- Naimi, A. I., & Balzer, L. B. (2018). Stacked generalization: An introduction to super learning. *European Journal of Epidemiology*, 33(5), 459–464. <https://doi.org/10.1007/s10654-018-0390-z>
- Nelder, J. A., & Wedderburn, R. W. M. (1972). Generalized linear models. *Journal of the Royal Statistical Society: Series A (General)*, 135(3), 370–384. <https://doi.org/10.2307/2344614>
- Pan, Y., & Zhao, Y.-Q. (2021). Improved doubly robust estimation in learning optimal individualized treatment rules. *Journal of the American Statistical Association*, 116(533), 283–294. <https://doi.org/10.1080/01621459.2020.1725522>
- Pedroza, C., & Truong, V. T. (2016). Performance of models for estimating absolute risk difference in multicenter trials with binary outcome. *BMC Medical Research Methodology*, 16(1), 113. <https://doi.org/10.1186/s12874-016-0217-0>
- Polley, E. (2018). *Super learner prediction*. Retrieved from <https://mran.microsoft.com/snapshot/2018-03-30/web/packages/SuperLearner/SuperLearner.pdf>
- Polley, E., Rose, S., & van der Laan, M. J. (2011). Super learning. In S. Rose & M. J. van der Laan (Eds.), *Targeted Learning* (pp. 43–66). Springer. https://doi.org/10.1007/978-1-4419-9782-1_3
- Prokhorenkova, L., Gusev, G., Vorobev, A., Dorogush, A. V., & Gulin, A. (2019). *CatBoost: Unbiased boosting with categorical features*. Retrieved from <https://papers.nips.cc/paper/2018/file/14491b756b3a51daac41c24863285549-Paper.pdf>
- Qiu, H., Carone, M., Sadikova, E., Petukhova, M., Kessler, R. C., & Luedtke, A. (2021). Optimal individualized decision rules using Instrumental Variable Methods. *Journal of the American Statistical Association*, 116(533), 174–191. <https://doi.org/10.1080/01621459.2020.1745814>
- R Core Team. (2020). *R: A language and environment for statistical computing*. The R Foundation.
- Reale, C., Novak, L. L., Robinson, K., Simpson, C. L., Ribeiro, J. D., Franklin, J. C., Ripberger, M., & Walsh, C. G. (2021). User-centered design of a Machine Learning Intervention for suicide risk prediction in a military setting. *AMIA Annual Symposium Proceedings*, 2020, 1050–1058.
- Robertson, S. E., Leith, A., Schmid, C. H., & Dahabreh, I. J. (2020). Assessing heterogeneity of treatment effects in observational studies. *American Journal of Epidemiology*, 190(6), 1088–1100. <https://doi.org/10.1093/aje/kwaa235>
- Robins, J. M., & Rotnitzky, A. (1995). Semiparametric efficiency in multi-variate regression models with missing data. *Journal of the American Statistical Association*, 90(429), 122–129. <https://doi.org/10.1080/01621459.1995.10476494>
- Rose, S., & van der Laan, M. J. (2014). A double robust approach to causal effects in case-control studies. *American Journal of Epidemiology*, 179(6), 663–669. <https://doi.org/10.1093/aje/kwt318>
- Rossom, R. C., Richards, J. E., Sterling, S., Ahmedani, B., Boggs, J. M., Yarborough, B. J. H., Beck, A., Lloyd, K., Frank, C., Liu, V., Clinch, S. B., Patke, L. D., & Simon, G. E. (2021). Connecting research and practice: Implementation of suicide prevention strategies in learning health care systems. *Psychiatric Services*, Advance online publication. <https://doi.org/10.1176/appi.ps.202000596>
- Rubin, D. B. (2008). For objective causal inference, design trumps analysis. *Annals of Applied Statistics*, 2(3), 808–840. <https://doi.org/10.1214/08-AOAS187>
- Ryan, A. T., Postolache, T. T., Taub, D. D., Wilcox, H. C., Ghahramanlou-Holloway, M., Umhau, J. C., & Deuster, P. A. (2021). Serum fatty acid latent classes are associated with suicide in a large military personnel sample. *Journal of Clinical Psychiatry*, 82(2). <https://doi.org/10.4088/JCP.20m13275>
- Salazar de Pablo, G., Studerus, E., Vaquerizo-Serrano, J., Irving, J., Catalan, A., Oliver, D., Baldwin, H., Danese, A., Fazel, S., Steyerberg, E. W., Stahl, D., & Fusar-Poli, P. (2021). Implementing precision psychiatry: A systematic review of individualized prediction models for clinical

- practice. *Schizophrenia Bulletin*, 47(2), 284–297. <https://doi.org/10.1093/schbul/sbaa120>
- SAS Institute Inc. (2013). *SAS®9.4 Software*. SAS Institute Inc.
- Singh, P. (2018). Treatment of vitamin D deficiency and comorbidities: A review. *Journal of the Association of Physicians of India*, 66(1), 75–82.
- Sun, J. X., Sinha, S., Wang, S. & Maiti, T. (2011). Bias reduction in conditional logistic regression. *Statistics in Medicine*, 30(4), 348–355. <https://doi.org/10.1002/sim.4105>
- Swanson, S. A. (2017). Instrumental variable analyses in pharmacoepidemiology: What target trials do we emulate? *Current Epidemiology Reports*, 4(4), 281–287. <https://doi.org/10.1007/s40471-017-0120-1>
- Tibshirani, R. (1996). Regression shrinkage and selection via the Lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(1), 267–288.
- Umhau, J. C., George, D. T., Heaney, R. P., Lewis, M. D., Ursano, R. J., Heilig, M., Hibbeln, J. R., & Schwandt, M. L. (2013). Low vitamin D status and suicide: A case-control study of active duty military service members. *PLoS One*, 8(1), e51543. <https://doi.org/10.1371/journal.pone.0051543>
- VA Office of Public and Intergovernmental Affairs. (2017). *VA REACH VET initiative helps save veterans lives: Program signals when more help is needed for at-risk veterans*. Retrieved from <https://www.va.gov/opa/pressrel/pressrelease.cfm?id=2878>
- Van Calster, B., Wynants, L., Verbeek, J. F. M., Verbakel, J. Y., Christodoulou, E., Vickers, A. J., Roobol, M. J., & Steyerberg, E. W. (2018). Reporting and interpreting decision curve analysis: A guide for investigators. *European Urology*, 74(6), 796–804. <https://doi.org/10.1016/j.eururo.2018.08.038>
- van der Laan, M. J., & Luedtke, A. R. (2015). Targeted learning of the mean outcome under an optimal dynamic treatment rule. *Journal of Causal Inference*, 3(1), 61–95. <https://doi.org/10.1515/jci-2013-0022>
- van der Laan, M. J., Polley, E. C., & Hubbard, A. E. (2007). Super learner. *Statistical Applications in Genetics and Molecular Biology*, 6(1), 25. <https://doi.org/10.2202/1544-6115.1309>
- van Klaveren, D., Balan, T. A., Steyerberg, E. W., & Kent, D. M. (2019). Models with interactions overestimated heterogeneity of treatment effects and were prone to treatment mistargeting. *Journal of Clinical Epidemiology*, 114, 72–83. <https://doi.org/10.1016/j.jclinepi.2019.05.029>
- VanderWeele, T. J., Luedtke, A. R., van der Laan, M. J., & Kessler, R. C. (2019). Selecting optimal subgroups for treatment using many covariates. *Epidemiology*, 30(3), 334–341. <https://doi.org/10.1097/ede.0000000000000991>
- Venkataramani, A. S., Bor, J., & Jena, A. B. (2016). Regression discontinuity designs in healthcare research. *BMJ*, 352, i1216. <https://doi.org/10.1136/bmj.i1216>
- Visconti, G., & Zubizarreta, J. R. (2018). Handling limited overlap in observational studies with cardinality matching. *Observational Studies*, 4(1), 217–249.
- Wager, S., & Athey, S. (2018). Estimation and inference of heterogeneous treatment effects using Random Forests. *Journal of the American Statistical Association*, 113(523), 1228–1242. <https://doi.org/10.1080/01621459.2017.1319839>
- Wang, Y., Fu, H., & Zeng, D. (2018). Learning optimal personalized treatment Rules in consideration of benefit and risk: With an application to treating type 2 diabetes patients with insulin therapies. *Journal of the American Statistical Association*, 113(521), 1–13. <https://doi.org/10.1080/01621459.2017.1303386>
- Wright, M. N., & Ziegler, A. (2017). ranger: A fast implementation of random forests for High Dimensional Data in C++ and R. *Journal of Statistical Software*, 77(1), 1–17. <https://doi.org/10.18637/jss.v077.i01>
- Zalsman, G., Hawton, K., Wasserman, D., van Heeringen, K., Arensman, E., Sarchiapone, M., Carli, V., Höschl, C., Barzilay, R., Balazs, J., Purebl, G., Kahn, J. P., Sáiz, P. A., Lipsicas, C. B., Bobes, J., Cozman, D., Hegerl, U., & Zohar, J. (2016). Suicide prevention strategies revisited: 10-year systematic review. *The Lancet Psychiatry*, 3(7), 646–659. [https://doi.org/10.1016/S2215-0366\(16\)30030-X](https://doi.org/10.1016/S2215-0366(16)30030-X)
- Zubizarreta, J. R., Small, D. S., Goyal, N. K., Lorch, S., & Rosenbaum, P. R. (2013). Stronger instruments via integer programming in an observational study of late preterm birth outcomes. *Annals of Applied Statistics*, 7(1), 25–50. <https://doi.org/10.1214/12-AOAS582>

How to cite this article: Zubizarreta, J. R., Umhau, J. C., Deuster, P. A., Brenner, L. A., King, A. J., Petukhova, M. V., Sampson, N. A., Tizenberg, B., Upadhyaya, S. K., RachBeisel, J. A., Streeten, E. A., Kessler, R. C., & Postolache, T. T. (2022). Evaluating the heterogeneous effect of a modifiable risk factor on suicide: The case of vitamin D deficiency. *International Journal of Methods in Psychiatric Research*, 31(1), e1897. <https://doi.org/10.1002/mpr.1897>

APPENDIX A

TABLE A1 Lasso on the training subsample of deployed matched pairs ($N = 328$ observations, 164 matched pairs)

	ODDS RATIO
Intercept	1.1
Main effects	
Race and ethnicity	
Identified as Black on race and ethnicity variable	1.0
Race: American Indian/Alaskan Native	2.1
Military	
Rank: Officer	0.9
History of military deployment coded in MH encounter	0.8
Total months on military deployment	1.0
Mental health	
Alcohol use disorder not otherwise specified	0.8
Number of DOD Inpatient mental health encounters	1.0
Number of non-Personality disorder MH diagnoses in record	1.1
Other, mixed, or unspecified drug abuse, unspecified	0.8
Number of inpatient mental health encounters	1.1
Number of mental health encounters in the 30 days preceding suicide	1.1
Any encounters for occupational therapy	1.1
Obesity, unspecified	1.4
Any mental health visits	1.4
Any Personality disorder	1.1
Biomarkers	
Docosapentaenoic acid (DPA; 22:5 n-6) $\mu\text{g}/\text{cl}$ in serum	1.0
Stearic acid or octadecanoic acid (18:0) as Percent of total fatty acids	0.2
Standard score of stearic acid (18:0) as Percent of total fatty acids	1.0
Concentration of this Palmitoleic acid expressed as a percentage of total fatty acid concentration expressed as a Z-score	1.0
Palmitoleic acid (16:1 n-7) as Percent of total fatty acids	
Activity of delta 9 desaturase (ratio of Palmitoleic acid and palmitic)	1.0
FA cluster: Risky versus protective	2.8
Dihomo- γ -linolenic acid (DGLA; 20:3 n-6) as Percent of total fatty acids (Percent of DGLA)	0.9
Ratio of stearic acid to palmitic acid	0.9
Magnesium $\mu\text{g}/\text{ml}$ in serum (mg)	1.0
Zinc $\mu\text{g}/\text{ml}$ in serum	0.9

Abbreviations: DGLA, dihomogamma-linolenic acid; FA, fatty acid; OR, odds ratio.