


ORIGINAL ARTICLE

Developing a practical suicide risk prediction model for targeting high-risk patients in the Veterans health Administration

Ronald C. Kessler¹  | Irving Hwang¹ | Claire A. Hoffmire² | John F. McCarthy³ | Maria V. Petukhova¹ | Anthony J. Rosellini⁴ | Nancy A. Sampson¹ | Alexandra L. Schneider² | Paul A. Bradley⁵ | Ira R. Katz⁶ | Caitlin Thompson^{7,8} | Robert M. Bossarte^{9,10}

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

²VISN 19 Mental Illness Research, Education and Clinical Care Center, Denver, Colorado, USA

³Office of Mental Health Operations, VA Center for Clinical Management Research, Serious Mental Illness Treatment Resource and Evaluation Center, Ann Arbor, Michigan, USA

⁴Center for Anxiety and Related Disorders, Boston University, Boston, Massachusetts, USA

⁵PricewaterhouseCoopers PS LLP, Washington, District of Columbia, USA

⁶Office of Mental Health Operations, Veterans Health Administration, Washington, District of Columbia, USA

⁷Office of Suicide Prevention, Veterans Health Administration, Washington, District of Columbia, USA

⁸Department of Psychiatry, University of Rochester, Rochester, New York, USA

⁹West Virginia University Injury Control Research Center and Department of Behavioral Medicine and Psychiatry, West Virginia University School of Medicine, Morgantown, West Virginia, USA

¹⁰Office of Suicide Prevention and VISN 2 Center of Excellence for Suicide Prevention, Veterans Health Administration, Washington, District of Columbia, USA

Correspondence

Ronald C. Kessler, Department of Health Care Policy, Harvard Medical School, 180A Longwood Avenue, Boston, Massachusetts 02115, USA.

Email: kessler@hcp.med.harvard.edu

Funding information

Department of Veterans Affairs Center for Innovation, Grant/Award Number: VA118-14-C-0046

Abstract

Objectives: The US Veterans Health Administration (VHA) has begun using predictive modeling to identify Veterans at high suicide risk to target care. Initial analyses are reported here.

Methods: A penalized logistic regression model was compared with an earlier proof-of-concept logistic model. Exploratory analyses then considered commonly-used machine learning algorithms. Analyses were based on electronic medical records for all 6,360 individuals classified in the National Death Index as having died by suicide in fiscal years 2009–2011 who used VHA services the year of their death or prior year and a 1% probability sample of time-matched VHA service users alive at the index date ($n = 2,112,008$).

Results: A penalized logistic model with 61 predictors had sensitivity comparable to the proof-of-concept model (which had 381 predictors) at target thresholds. The machine learning algorithms had relatively similar sensitivities, the highest being for Bayesian additive regression trees, with 10.7% of suicides occurred among the 1.0% of Veterans with highest predicted risk and 28.1% among the 5.0% of with highest predicted risk.

Conclusions: Based on these results, VHA is using penalized logistic regression in initial intervention implementation. The paper concludes with a discussion of other practical issues that might be explored to increase model performance.

KEYWORDS

assessment/diagnosis, clinical decision support, epidemiology, machine learning, predictive modeling, suicide/self harm

1 | INTRODUCTION

The US Department of Veteran Affairs (VA) is the cabinet-level department with responsibility for providing services and benefits to US Military Veterans. The VA has three main subdivisions: the Veterans Benefits Administration (compensation and pensions, home loans, insurance, vocational services, education through the GI Bill); the Veterans Health Administration (VHA; health care and biomedical research); and the National Cemetery Administration (burial services and maintenance of VA cemeteries). VHA is the largest of these subdivisions and also the largest integrated health care system in the United States, with 168 VA Medical Centers and 1,053 outpatient clinics that currently serve over six million Veterans each year (<https://www.va.gov/health/>).

The most recent estimates suggest that an average of 20 Veterans die by suicide each day in the United States, representing 18% of all US suicide deaths among individuals ages 18 and older, which is substantially higher than expected given that Veterans make up 8.5% of the population (Office of Suicide Prevention in Veterans Health Administration, 2016). Six of these deaths occur among current and recent users of VHA health care services. A new VHA program addresses this problem using a statistical prediction model to target Veterans using VHA services deemed to be at highest suicide risk for a preventive intervention (Office of Public and Intergovernmental Affairs in Veterans Health Administration, 2017). The feasibility of using such a model was demonstrated in a proof-of-concept study by McCarthy et al. (2015), which showed that a logistic regression model using VHA data could significantly predict future suicides. However, multicollinearity among the predictors in that model raised concerns that prediction accuracy might be lower than in a model containing fewer predictors. The current report presents the results of an analysis designed to improve on the McCarthy model using the same logistic link function and initial predictor set but selecting a smaller set of predictors. We also explored the possibility that more complex algorithms might improve prediction accuracy. The paper closes with a discussion of important practical considerations for future modeling and program planning.

2 | MATERIALS AND METHODS

2.1 | Sample

We began with the same database as in the McCarthy et al. (2015) analysis, which consisted of all 6,360 individuals classified in the National Death Index (NDI; Centers for Disease Control and Prevention & Department of Health and Human Services, 2015) as having died by suicide in fiscal years 2009–2011 (October 1, 2008–September 30, 2011) who used VHA services in the year of their death or the prior year and a 1% probability sample of time-matched (to suicide decedents) patients alive at the end of the month the suicide decedent died who received VHA services over the same period of time ($n = 2,112,008$). The logic of the data array was that of discrete-time survival analysis with person-month the unit of analysis and time-varying predictors defined as of the month before the death

(Willett & Singer, 1993). The controls received a weight of 100 (i.e. 1/1.0%) to adjust for the under-sampling of non-cases, which was implemented to reduce computational intensity. An average of 176.7 recorded suicides occurred per month in this population over the study period, equivalent to 36.1 per 100,000 person-years among the roughly 5.9 million Veterans meeting study criteria at a point in time. Unlike McCarthy et al. (2015), we excluded the 29 original sample members for whom administrative data were missing on patient gender or age as well as the 3,484 original sample members who were classified as either younger than 18 or older than 100 at the date of death, resulting in a final sample of 6,359 cases and 2,108,496 controls.

McCarthy et al. (2015) divided the sample into random halves, estimated coefficients in one half, then applied these coefficients to the other half to check for out-of-sample model performance, and then created a prospective sample of all individuals who were alive as of September 30, 2010 and had received VHA services in fiscal year 2010 or 2011 to explore other outcomes for the patients who were identified as being at high risk, including deaths from suicide over a 12-month time horizon. It is noteworthy that 33% of the suicide decedents in the model development sample (fiscal years 2009–2011) were also included in the prediction sample (fiscal year 2011), resulting in lack of independence. We used a different approach to validate our model that corrected this problem by dividing suicide decedents in fiscal years 2009–2010 and their controls into random halves to create separate training and contemporaneous test samples and then applying these coefficients to an independent prospective fiscal year 2011 validation sample. We used a consistent 30-day time horizon both in estimating and evaluating model fit, again in order to be consistent with McCarthy et al. (2015), even though an argument could be made for alternative time horizons being of equal or greater clinical and policy importance. We return to the issue of alternative time horizons in the discussion section.

2.2 | Predictors

In order to facilitate direct comparison, we considered the same predictors as McCarthy et al. (2015): 381 measures of VHA service use as defined over the 730 days before the death (or selection as a control). As described by McCarthy et al. (2015), these predictors were selected based on evidence in previous empirical studies of risk factors for suicide and on the availability of appropriate indicators in VHA electronic medical records. Given our focus on overall model prediction accuracy across different estimation methods rather than substantive interpretation of individual predictors, and given the large number of predictors considered by McCarthy et al. (2015), we do not provide details about these predictors here but only note that they assessed variables in five broad domains that have been shown in previous research to predict future suicides (Kessler et al., 2015; Kessler, Stein et al., 2017): intensity-recency of inpatient, outpatient, and emergency service use for various mental disorders over time lags between 1 and 24 months; parallel measures of VHA service use for other health problems; measures of prescriptions filled for various classes of psychotropic and other medications over the same time periods; basic socio-demographic variables (e.g. age, sex, region

of the country); and a number of interactions between socio-demographics and selected health care measures. Interested readers are referred to the original McCarthy et al. (2015) paper for a more detailed description of the precise predictors.

2.3 | Analysis methods

De-identified data analysis was carried out remotely on a secure VHA server by Harvard Medical School analysts with approval by the Harvard Medical School Human Subjects Committee. Model-building began by estimating the McCarthy et al. (2015) model in the training sample and applying coefficients to the test and prospective validation samples to evaluate out-of-sample performance in determining the proportions of suicides among the 0.1%, 1.0%, and 5.0% of VHA patients with the highest predicted probabilities of suicide death (thresholds used by McCarthy et al., 2015). It is conventional in such analyses to evaluate prediction accuracy at each threshold by estimating the test operating characteristics of sensitivity (the proportion of suicides among Veterans with predicted probabilities above the threshold), specificity (the proportion of non-suicides among Veterans with predicted probabilities below the threshold), positive predictive value (PPV; the proportions of screened positives that did, in fact, commit suicide), negative predictive value (NPV; the proportion of screened negative that did not commit suicide), and area under the receiver operating characteristic curve (AUC; the probability that a randomly selected true case had a higher predicted probability than a randomly selected non-case). However, given the rarity of death from suicide, we focus here only on sensitivity, as specificity and NPV will be very close to 1 – the threshold regardless of sensitivity and PPV will be no higher than 0.3% (i.e. 99.7% of screened positives would not commit suicide over a 30-day time horizon) even if 100% of true suicide deaths occurred among Veterans above the 0.1% threshold. The feasibility of developing interventions for such a rare outcome is a separate matter considered in the discussion section.

As noted earlier in describing the sample, McCarthy et al. (2015) included all suicide deaths and a 1% sample of other person-months in the sample. This kind of under-sampling of non-cases is one of the standard approaches used to deal with the problem of “class imbalance,” which occurs when the outcome of interest is rare (He & Garcia, 2009). The problem here is that most prediction algorithms aim to optimize overall classification accuracy and fail to adjust for the fact that false negatives may be more costly than false positives, leading the algorithms to focus on correctly classifying the much more common non-cases at the cost of misclassifying the rare cases. A number of strategies involving under-sampling of non-cases, pseudo-sampling of cases, and combinations have been developed to address this problem (Chawla, 2010). Some of these approaches have been shown to improve on simple sub-sampling (e.g. Lee, 2014; Rahman & Davis, 2013). However, in order to maintain comparability with the McCarthy et al. (2015) analysis, we retained their sampling design in our analysis rather than use alternative approaches to address the problem of the highly skewed outcome distribution.

The McCarthy et al. (2015) model, which included all 381 predictors, was estimated with *proc logistic* in SAS 9.3 (SAS Institute Inc, 2010). However, this model was under-identified due to perfect multivariate associations among some model predictors. This identification problem was resolved in SAS by the program excluding the redundant predictors to achieve convergence, but this kind of over-fitting is known to reduce out-of-sample performance (Upstill-Goddard, Eccles, Fliege, & Collins, 2013). The challenge in refining the model was to select an optimal subset of predictors to avoid over-fitting. We did this initially by using elastic net penalized regression (Zou & Hastie, 2005) estimated with the R-package *glmnet* (Friedman, Hastie, & Tibshirani, 2010) to select the best additive subset of predictors to optimize classification of future suicide deaths. Elastic net regression penalizes over-fitting with a composite penalty that combines a ridge penalty (which handles multicollinearity by shrinking all coefficients smoothly towards zero but retains all variables in the model) (Hoerl & Kennard, 1970) and a lasso penalty (which 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) (Tibshirani, 1996). A range of elastic net models that varied the relative importance of the two penalties was estimated in the training sample and applied in the test sample to decide on an optimal mix. This elastic net approach of combining the ridge and lasso penalties has the advantage of yielding more stable and accurate estimates than either alone while maintaining model parsimony and using the same link function (i.e. logistic model assuming additivity among predictors) as the original McCarthy et al. (2015) model (Zou & Hastie, 2005). Estimates of sensitivity based on the final elastic net model and the McCarthy et al. (2015) model were compared in the independent prospective validation sample among the 0.1%, 1.0%, and 5.0% of Veterans with highest predicted probabilities of suicide in each model.

We then investigated whether more complex machine learning models would yield higher sensitivities by working with eight machine learning algorithms that allow complex non-linearities and interactions among predictors. These algorithms, which were selected based on prior recommendations in the literature (Fernández-Delgado, Cernadas, Barro, & Amorim, 2014; Wu et al., 2008), included two decision tree algorithms (Bayesian additive regression trees [BART; Chipman, George, & McCulloch, 2010]; random forest [Breiman, 2001]), two spline algorithms (adaptive splines [Friedman, 1991]; adaptive polynomial splines [Stone, Hansen, Kooperberg, & Truong, 1997]), generalized boosting (Freund & Schapire, 1999), and support vector machines with linear, polynomial, and radial kernels (Steinwart & Christmann, 2008). A basic overview of each algorithm is provided in Table 1. Each algorithm was implemented in the training sample using internal cross-validation to select the optimal specification, tuned in the test sample to set optimal hyper-parameter values, and then applied in the independent prospective validation sample to compare out-of-sample performance with the McCarthy et al. (2015) and elastic net models. Importantly, out-of-sample performance was tested in data for future years in the independent prospective validation sample (i.e. fiscal year 2009–2010 data used to develop the models and fiscal year 2011 data used to evaluate model performance), as this is the way the final model will be used in practice by VHA in the future.

TABLE 1 Overview of the algorithms used in the analysis

Algorithm	R package	Description
I. Spline		<ul style="list-style-type: none"> • adaptive spline regression flexibly captures interactions and linear and non-linear associations
Adaptive splines	<i>earth</i> (Milborrow, Hastie, Tibshirani, Miller, & Lumley, 2016)	<ul style="list-style-type: none"> • linear segments (splines) of varying slopes are connected and smoothed to create piece-wise curves (basis functions) • final fit is built using a stepwise procedure that selects the optimal combination of basis functions
Adaptive polynomial splines	<i>polyspline</i> (Kooperberg, 2015)	<ul style="list-style-type: none"> • earth and polymars are generally similar, but differ in the order in which basis functions (e.g. linear versus non-linear) are added to build the final model
II. Decision tree		<ul style="list-style-type: none"> • decision tree methods capture interactions and non-linear associations
Random forest	<i>randomForest</i> (Liaw & Wiener, 2002)	<ul style="list-style-type: none"> • independent variables are partitioned (based on values) and stacked to build decision trees and ensemble an aggregate “forest” • random forest builds numerous trees in bootstrapped samples and generates an aggregate tree by averaging across trees (reducing overfit)
Bayesian additive regression trees (BART)	<i>BayesTree</i> (Chipman & McCulloch, 2016)	<ul style="list-style-type: none"> • Bayesian trees are based on an underlying probability model (priors) for the structure and likelihood for data in terminal nodes; aggregate tree is generated by averaging across tree posteriors (reducing overfit)
III. Support vector machines (SVM)	<i>e1401</i> (Meyer et al., 2015)	<ul style="list-style-type: none"> • support vector machines treat each independent variable as dimensions in high dimensional space and attempt to identify the best hyperplane to separate the sample into classes (e.g. cases and non-cases)
Linear kernel		<ul style="list-style-type: none"> • goal is to find the hyperplane with the maximum margin between the two closest points in space
Polynomial kernel		<ul style="list-style-type: none"> • captures linear associations, but alternate kernels can be used to capture non-linearities (polynomial and radial basis kernels were used here)
Radial kernel		
IV. Generalized boosted regression models		
Adaptive boosting	<i>gbm</i> (Freund & Schapire, 1999)	<ul style="list-style-type: none"> • adaptive boosting is a meta-algorithm that iteratively fits decision-trees using weights to adjust for cases classified incorrectly in the prior iteration • this allows subsequent iterations to focus on predicting more difficult cases

3 | RESULTS

3.1 | Performance of the optimal elastic net model compared to the McCarthy et al. (2015) model

The best elastic net model in the training sample had 61 predictors and exclusively used the lasso penalty. Sensitivities of that model in the test sample among patients in the top 0.1%, 1.0%, and 5.0% of risk were 2.8%, 11.8%, and 28.2%, respectively (Table 2). These values were all comparable to or slightly higher than those of the McCarthy et al. (2015) model even though the latter included 381 predictors. Sensitivities were lower in all models in the prospective validation sample but the sensitivities of the 61-variable elastic net model remained the same or higher (2.2–26.3%) than those of the 381-variable McCarthy et al. (2015) model (2.0–25.3%).

3.2 | Performance of machine learning models that allowed for non-linearities and interactions

A number of the other algorithms we considered required tuning in the test sample to fix parameters that had to be specified in advance for the model to converge. As we used the test sample for this purpose, we focus only on comparative model performance in the prospective validation sample. BART had the highest sensitivity among the 0.1% of patients with highest predicted risk (2.7%) followed by adaptive polynomial regression splines (2.4%) and elastic net (2.2%) (Table 3). BART also had highest sensitivity among the 1% of patients with

TABLE 2 Comparative model fit of best-fitting elastic net model with the McCarthy et al. (2015) model estimated in the fiscal years 2009–2010 training sample and applied to both the fiscal years 2009–2010 test sample and the independent prospective fiscal year 2011 validation sample

	Sensitivity among Veterans with predicted risks in the top ...		
	0.1%	1.0%	5.0%
I. Elastic net model applied to			
Testing sample	2.8	11.8	28.2
Validation sample	2.2	9.9	26.3
II. McCarthy model applied to			
Testing sample	2.9	11.6	27.1
Validation sample	2.0	9.5	25.3

highest predicted risk (10.7%) followed by elastic net (9.9%) and gradient boosting (9.8%). BART again had highest sensitivity among the 5% of patients with highest predicted risk (28.1%) followed by gradient boosting (27.0%) and elastic net (26.3%).

4 | DISCUSSION

We showed that a penalized logistic model containing only 61 predictors had comparable sensitivity in an independent prospective validation sample to the logistic model with 381 predictors in the original McCarthy et al. (2015) analysis. We also showed that more complex

TABLE 3 Comparative model fit of elastic net with other machine learning classifiers estimated in the fiscal years 2009–2010 training sample, tuned in fiscal years 2009–2010 test sample, and applied in the independent prospective fiscal year 2011 validation sample

	Sensitivity among Veterans with predicted risks in the top ...		
	0.1%	1.0%	5.0%
I. Elastic net	2.2	9.9	26.3
II. Splines			
Adaptive splines	1.8	9.0	24.0
Adaptive polynomial splines	2.4	9.6	26.0
III. Decision trees			
Random forest	2.0	9.3	24.1
Bayesian additive regression trees (BART)	2.7	10.7	28.1
IV. Support vector machines (SVM)			
Linear kernel	1.0	6.1	17.3
Polynomial kernel	1.0	7.1	19.7
Radial kernel	1.2	6.9	21.2
V. Generalized adaptive boosting	2.0	9.8	27.0

machine learning algorithms allowing for non-linearities and interactions had comparable sensitivities at the same thresholds, but with BART seeming to have a slight advantage over the other algorithms. BART uses Bayesian averaging of regression trees across multiple samples to address the problem of over-fitting that exists in random forests and other tree-based approaches. Comparison studies have shown that BART often out-performs other commonly-used machine learning algorithms, including random forests, neural networks, and gradient boosting, in head-to-head comparisons (Chipman et al., 2010). However, to confirm the reliability of this advantage in predicting Veteran suicides, it would be useful to evaluate the stability of the relatively modest advantage we found here for BART by carrying out simulations to calculate the standard errors of the sensitivity estimates and replicating the analyses over different years and time lags. Based on the relatively modest advantages of BART and the other complex machine learning methods over penalized logistic regression in the analyses reported here, VHA is using the penalized logistic model to target Veterans for preventive intervention while the possibility of using more complex models is under investigation.

Taken together with recent advances from the literature, the findings presented here suggest a number of opportunities for enhancing and extending the current model. First, alternative methods could help deal with the problem of extreme imbalance (i.e. the rarity of suicide deaths). As noted earlier in the section on analysis methods, a number of methods have been developed to address this problem (Chawla, 2010). Toolkits exist to evaluate the relative effectiveness of these different methods in specific empirical cases (Kuhn, 2015; Lemaitre, Nogueira, & Aridas, 2016). We are carrying out a systematic comparison of these different methods to determine the best one for predicting VA suicide deaths.

Second, we are exploring the value of expanding the predictor set beyond information about treatment available in the VHA electronic medical record. Under consideration here are such things as: (i) residential zip code data to code small area geocode information on

variables known to predict suicides (e.g. local unemployment rate (Nordt, Warnke, Seifritz, & Kawohl, 2015); (ii) historical US Department of Defense administrative data known to predict post-discharge suicides (Reger et al., 2015); (iii) data from commercial search engines calling up various public records (e.g. legal, financial, criminal justice) that might predict suicides (<http://www.accurint.com/>); (iv) surveillance of postings on social media for patients who consent to monitoring; (v) surveillance of data from phone apps (Onnela & Rauch, 2016) and wearables (Alam, Cho, Huh, & Hong, 2014).

Third, we are exploring the possibility that prediction accuracy could be improved not only by using machine learning methods that allow for complex non-linear-interactive associations, but also by combining predictions across algorithms rather than selecting one best algorithm. This *ensemble* approach can be especially useful when certain algorithms predict some types of cases better than others. For example, the SuperLearner ensemble method yields a level of prediction accuracy at least as high as that of the best-performing algorithm in the ensemble set and often considerably higher than that value (Polley, LeDell, Kennedy, Lendle, & van der Laan, 2016). As a result, the questions that need to be investigated are which algorithms to include in the ensemble and whether the level of improvement in prediction accuracy based on the ensemble compared to the best single algorithm is sufficient to warrant the increased effort of using the ensemble approach. We are exploring both of these issues.

Fourth, the 30-day time horizon used by McCarthy et al. (2015) (which, as noted in the section on the sample, we accepted for purposes of comparison) needs to be reconsidered. In characterizing the individuals identified as being at high risk in their model, McCarthy et al. (2015) found that they were at increased risk for a period of at least one year. However, the risk decayed rapidly over longer time horizons, especially for the highest-risk patients. This should not be surprising, as the optimal predictors of imminent suicide risk are unlikely to be the same as the predictors of suicides over a longer time period. The only way to address this issue, recognizing that longer time horizons are of clinical and policy importance, is to estimate models that allow for different predictors (or different coefficients associated with the same predictors) for different time horizons. One approach of this sort would be to use survival analysis to predict suicide deaths over a longer time horizon than the 30 days considered here and build into the model the possibility that some baseline predictors vary in their coefficients with increased time from baseline (van Houwelingen & Putter, 2011). Another approach would be to estimate separate models for different time windows (e.g. suicide deaths in the 30 days from baseline, in days 31–60 from baseline, days 61–90 from baseline, etc.), and evaluate the extent to which prediction accuracy decays over time. Both approaches are promising.

Fifth, the possibility is being investigated of developing models to predict which high-risk patients are most likely to be helped by specific interventions to complement models that predict which patients have the highest suicide risks (Kessler, van Loo et al., 2017). The two types of predictions need not be the same, as patients at the very highest risk might be less responsive than those at slightly lower risk to some preventive interventions. Consistent with this possibility, clinical researchers have shown that mentally ill patients differ not only in *absolute* treatment response (i.e. the impact

of treatment on a given patient) but also in *relative* treatment response (i.e. the specific treatment that is optimal for a given patient) and that a wide range of variables other than disorder severity predicts both types of differences (Kessler, van Loo et al., 2017).

One way to advance our understanding of differential treatment response would be for the VHA to randomize their initial preventive intervention over a wider range of risk rather than implement the intervention only with the highest-risk patients (e.g. to intervene with a random one-tenth of the patients at the highest 1.0% of predicted risk rather than with all of the patients at highest 0.1% of predicted risk). This design would make it possible to search for systematic predictors of differential treatment response using recently-developed machine learning methods developed for that purpose (Imai & Ratkovic, 2013; Rosenblum & van der Laan, 2011) and then to use the results to target future intervention assignments to the patients most likely to be helped and randomize additional interventions among patients less likely to be helped by the earlier interventions. This use of sequential pragmatic trials would make it possible to build an increasingly sophisticated clinical decision support scheme for optimizing patient treatment response across a coordinated set of interventions.

5 | CONCLUSIONS

Based on the results reported here, VHA has implemented a program using the elastic net model reported here to target patients for preventive interventions. At the same time, as a part of ongoing program development, VHA is considering the expansion of predictors to consider in future models and evaluating the extent to which more advanced machine learning algorithms and ensemble methods could improve prediction. It is evaluating the impact of developing models that use different approaches to address the class imbalance problems and that are designed specifically to allow prediction across the range of time horizons that are of importance to policy and practice. VHA is also strongly encouraging research to consider the benefits of strategies that target patients with the highest probabilities of responding to interventions rather than focusing only on those with the highest probabilities of death from suicide.

DECLARATION OF INTEREST STATEMENT

In the past three years, Dr Kessler received support for his epidemiological studies from Sanofi Aventis; was a consultant for Johnson & Johnson Wellness and Prevention, Shire, Takeda; and served on an advisory board for the Johnson & Johnson Services Inc. Lake Nona Life Project. Dr Kessler is a co-owner of DataStat, Inc., a market research firm that carries out healthcare research. Mr Bradley is an employee of PricewaterhouseCoopers. The remaining authors have no conflicts of interest to declare.

DISCLAIMER

The views and opinions expressed in this article are those of the authors and should not be construed to represent the views of any of the sponsoring organizations or agencies.

REFERENCES

- Alam, M. G. R., Cho, E. J., Huh, E. N., & Hong, C. S. (2014). Cloud based mental state monitoring system for suicide risk reconnaissance using wearable bio-sensors. In *Proceedings of the 8th International Conference on Ubiquitous Information Management and Communication* (Article No. 56). Siem Reap, Cambodia: ACM.
- Breiman, L. (2001). Random forests. *Machine Learning*, 45(1), 5–32. <https://doi.org/10.1023/A:1010933404324>
- Centers for Disease Control and Prevention & Department of Health and Human Services. (2015). National Death Index. <https://www.healthdata.gov/dataset/national-death-index> [1 August 2016]
- Chawla, N. V. (2010). Data mining for imbalanced datasets: An overview. In O. Maimon, & L. Rokach (Eds.), *Data Mining and Knowledge Discovery Handbook* (2nd ed.) (pp. 875–886). Berlin/Heidelberg: Springer.
- Chipman, H. A., George, E. I., & McCulloch, R. E. (2010). BART: Bayesian additive regression trees. *The Annals of Applied Statistics*, 4(1), 266–298. <https://doi.org/10.1214/09-AOAS285>
- Chipman, H. A., & McCulloch, R. E. (2016). BayesTree: Bayesian additive regression trees [computer program]. R package version 0.3–1.4. <http://CRAN.R-project.org/package=BayesTree>
- Fernández-Delgado, M., Cernadas, E., Barro, S., & Amorim, D. (2014). Do we need hundreds of classifiers to solve real world classification problems. *Journal of Machine Learning Research*, 15(1), 3133–3181. <http://www.jmlr.org/papers/volume15/delgado14a/source/delgado14a.pdf>
- Freund, Y., & Schapire, R. (1999). A short introduction to boosting. *Journal of Japanese Society for Artificial Intelligence*, 14(5), 771–780. (in Japanese, translation by Naoki Abe). <http://www.yorku.ca/gisweb/eats4400/boost.pdf>
- Friedman, J., Hastie, T., & Tibshirani, R. (2010). Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software*, 33(1), 1–22. <https://doi.org/10.18637/jss.v033.i01>
- Friedman, J. H. (1991). Multivariate adaptive regression splines. *The Annals of Statistics*, 19(1), 1–141. <http://www.jstor.org/stable/2241837>
- He, H., & Garcia, E. A. (2009). Learning from imbalanced data. *IEEE Transactions on Knowledge and Data Engineering*, 21(9), 1263–1284. <https://doi.org/10.1109/TKDE.2008.239>
- Hoerl, A. E., & Kennard, R. W. (1970). Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, 12(1), 55–67. <https://doi.org/10.2307/1267351>
- Imai, K., & Ratkovic, M. (2013). Estimating treatment effect heterogeneity in randomized program evaluation. *The Annals of Applied Statistics*, 7(1), 443–470. <https://doi.org/10.1214/12-AOAS593>
- Kessler, R. C., Stein, M. B., Petukhova, M. V., Bliese, P., Bossarte, R. M., Bromet, E. J., ... Ursano, R. J. (2017). Predicting suicides after outpatient mental health visits in the Army Study to Assess risk and Resilience in Servicemembers (Army STARRS). *Molecular Psychiatry*, 22(4), 544–551. <https://doi.org/10.1038/mp.2016.110>
- Kessler, R. C., van Loo, H. M., Wardenaar, K. J., Bossarte, R. M., Brenner, L. A., Ebert, D. D., ... Zaslavsky, A. M. (2017). Using patient self-reports to study heterogeneity of treatment effects in major depressive disorder. *Epidemiology and Psychiatric Sciences*, 26(1), 22–36. <https://doi.org/10.1017/s2045796016000020>
- Kessler, R. C., Warner, C. H., Ivany, C., Petukhova, M. V., Rose, S., Bromet, E. J., ... Ursano, R. J. (2015). Predicting suicides after psychiatric hospitalization in US Army soldiers: The Army Study to Assess Risk and Resilience in Servicemembers (Army STARRS). *JAMA Psychiatry*, 72(1), 49–57. <https://doi.org/10.1001/jamapsychiatry.2014.1754>
- Kooperberg, C. (2015). Pplspline: Polynomial spline routines [computer program]. R package version 1.1.12. <http://CRAN.R-project.org/package=pplspline>
- Kuhn, M. (2015). caret: Classification and regression training. Astrophysics Source Code Library, record ascl:1505.003 [computer program]. <http://adsabs.harvard.edu/abs/2015ascl.soft05003K>
- Lee, P. H. (2014). Resampling methods improve the predictive power of modeling in class-imbalanced datasets. *International Journal of*

- Environmental Research and Public Health*, 11(9), 9776–9789. <https://doi.org/10.3390/ijerph110909776>
- Lemaitre, G., Nogueira, F., & Aridas, C. K. (2016). Imbalanced-learn: A python toolbox to tackle the curse of imbalanced datasets in machine learning. *Journal of Machine Learning Research*, 7, 1–5. arXiv: 1609.06570
- Liaw, A., & Wiener, M. (2002). Classification and regression by randomForest. *R news*, 2(3), 18–22. <http://ai2-s2-pdfs.s3.amazonaws.com/6e63/3b41d93051375ef9135102d54fa097dc8cf8.pdf>
- McCarthy, J. F., Bossarte, R. M., Katz, I. R., Thompson, C., Kemp, J., Hannemann, C. M., ... Schoenbaum, M. (2015). Predictive modeling and concentration of the risk of suicide: Implications for preventive interventions in the US Department of Veterans Affairs. *American Journal of Public Health*, 105(9), 1935–1942. <https://doi.org/10.2105/AJPH.2015.302737>
- Meyer, D., Dimitriadou, E., Hornik, K., Weingessel, A., Leisch, F., Chang, C. C., & Lin, C. C. (2015). e1071: Misc Functions of the Department of Statistics, TU Wien [computer program]. R package version 1.5–7. <https://rdrr.io/rforge/e1071/>
- Milborrow, S., Hastie, T., Tibshirani, R., Miller, A., & Lumley, T. (2016). Earth: Multivariate adaptive regression splines [computer program]. R package version 4.4.5. <http://CRAN.R-project.org/package=earth>
- Nordt, C., Warnke, I., Seifritz, E., & Kawohl, W. (2015). Modelling suicide and unemployment: A longitudinal analysis covering 63 countries, 2000–11. *Lancet Psychiatry*, 2(3), 239–245. [https://doi.org/10.1016/S2215-0366\(14\)00118-7](https://doi.org/10.1016/S2215-0366(14)00118-7)
- Office of Public and Intergovernmental Affairs in Veterans Health Administration. (2017). VA REACH VET initiative helps save Veterans lives: Program signals when more help is needed for at-risk Veterans. <https://www.va.gov/opa/pressrel/pressrelease.cfm?id=2878> [12 May 2017]
- Office of Suicide Prevention in Veterans Health Administration. (2016). Suicide among Veterans and other Americans, 2001–2014. <https://www.mentalhealth.va.gov/docs/2016suicidedatareport.pdf> [1 July 2016]
- Onnela, J. P., & Rauch, S. L. (2016). Harnessing smartphone-based digital phenotyping to enhance behavioral and mental health. *Neuropsychopharmacology*, 41, 1691–1696. <https://doi.org/10.1038/npp.2016.7>
- Polley, E., LeDell, E., Kennedy, C., Lendle, S., & van der Laan, M. (2016). SuperLearner: Super learner prediction [computer program]. R package version 2.0–21: The Comprehensive R Archive Network. <https://cran.r-project.org/web/packages/SuperLearner/index.html>
- Rahman, M. M., & Davis, D. N. (2013). Addressing the class imbalance problem in medical datasets. *International Journal of Machine Learning and Computing*, 3(2), 224–228. <https://doi.org/10.7763/IJMLC.2013.V3.307>
- Reger, M. A., Smolenski, D. J., Skopp, N. A., Metzger-Abamukang, M. J., Kang, H. K., Bullman, T. A., ... Gahm, G. A. (2015). Risk of suicide among US Military service members following Operation enduring Freedom or Operation Iraqi Freedom deployment and separation from the US Military. *JAMA Psychiatry*, 72(6), 561–569. <https://doi.org/10.1001/jamapsychiatry.2014.3195>
- Rosenblum, M., & van der Laan, M. J. (2011). Optimizing randomized trial designs to distinguish which subpopulations benefit from treatment. *Biometrika*, 98(4), 845–860. <https://doi.org/10.1093/biomet/asr055>
- SAS Institute Inc. (2010). SAS/STAT® Software [computer program]. Version 9.3 for Unix. Cary, NC: SAS Institute Inc.
- Steinwart, I., & Christmann, A. (2008). *Support Vector Machines*. New York: Springer.
- Stone, C. J., Hansen, M., Kooperberg, C., & Truong, Y. K. (1997). 1994 Wald memorial lecture: Polynomial splines and their tensor products in extended linear modeling. *The Annals of Statistics*, 25(4), 1371–1470. <https://doi.org/10.1214/aos/1031594728>
- Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, 58(1), 267–288. <http://www.jstor.org/stable/2346178>
- Upstill-Goddard, R., Eccles, D., Fliege, J., & Collins, A. (2013). Machine learning approaches for the discovery of gene–gene interactions in disease data. *Briefings in Bioinformatics*, 14(2), 251–260. <https://doi.org/10.1093/bib/bbs024>
- van Houwelingen, H., & Putter, H. (2011). *Dynamic Prediction in Clinical Survival Analysis*. Boca Raton, FL: CRC Press.
- Willett, J. B., & Singer, J. D. (1993). Investigating onset, cessation, relapse, and recovery: Why you should, and how you can, use discrete-time survival analysis to examine event occurrence. *Journal of Consulting and Clinical Psychology*, 61(6), 952–965. <https://doi.org/10.1037/0022-006X.61.6.952>
- Wu, X., Kumar, V., Ross Quinlan, J., Ghosh, J., Yang, Q., Motoda, H., ... Steinberg, D. (2008). Top 10 algorithms in data mining. *Knowledge and Information Systems*, 14(1), 1–37. <https://doi.org/10.1007/s10115-007-0114-2>
- Zou, H., & Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodological)*, 67(2), 301–320. <https://doi.org/10.1111/j.1467-9868.2005.00503.x>

How to cite this article: Kessler RC, Hwang I, Hoffmire CA, et al. Developing a practical suicide risk prediction model for targeting high-risk patients in the Veterans health Administration. *Int J Methods Psychiatr Res*. 2017;26:e1575. <https://doi.org/10.1002/mpr.1575>