Clinical Psychology Review 82 (2020) 101940

Contents lists available at ScienceDirect

Clinical Psychology Review

journal homepage: www.elsevier.com/locate/clinpsychrev

Review

Machine learning for suicidology: A practical review of exploratory and hypothesis-driven approaches

Christopher R. Cox [a,*], Emma H. Moscardini [a], Alex S. Cohen [a,b], Raymond P. Tucker [a]

[a] Louisiana State University, Department of Psychology, USA
[b] Louisiana State University, Center for Computation and Technology, USA

ARTICLE INFO

Keywords:
Machine learning
Suicide
Suicidal thoughts and behaviors
Prediction
Structured sparsity

ABSTRACT

Machine learning is being used to discover models to predict the progression from suicidal ideation to action in clinical populations. While quantifiable improvements in prediction accuracy have been achieved over theory-driven efforts, models discovered through machine learning continue to fall short of clinical relevance. Thus, the value of machine learning for reaching this objective is hotly contested. We agree that machine learning, treated as a "black box" approach antithetical to theory-building, will not discover clinically relevant models of suicide. However, such models may be developed through deliberate synthesis of data- and theory-driven approaches. By providing an accessible overview of essential concepts and common methods, we highlight how generalizable models and scientific insight may be obtained by incorporating prior knowledge and expectations to machine learning research, drawing examples from suicidology. We then discuss challenges investigators will face when using machine learning to discover models of low prevalence outcomes, such as suicide.

1. Introduction

Despite decades of investigation into motives and risk factors, anticipating suicide remains a practical challenge (Franklin et al., 2017). Recently, state-of-the-art machine learning techniques have been applied in large clinical datasets to discover probabilistic relationships between patient data and psychiatric risk (Beam & Kohane, 2018; Bzdok & Meyer-Lindenberg, 2018) including suicidal thoughts and behaviors (STBs; Belsher et al., 2019). Exploratory data-driven machine learning can construct powerful prediction models that rely on relationships among variables that are unanticipated by existing theory and are too complicated or subtle to be detected by clinical observation or in smaller-scale experiments (Ribeiro et al., 2016). Although machine learning has already discovered models of suicide risk that outperform previous efforts, the approach may return complex models that are difficult to understand, gain insight from, or generalize to new situations. These concerns, as well as the utility of machine learning for predicting suicide risk in general, has been raised in the extant literature (Belsher et al., 2019; Cabitza, Rasoini, & Gensini, 2017; McHugh & Large, 2020).

The objective of the current work is not to review or critique applications of predictive modeling and machine learning in suicidology (cf. Cabitza et al., 2017). Rather, in light of the excitement (and consternation) garnered by recent applications of machine learning for suicide prediction, we argue that merely replacing a theory-driven approach to understanding suicide risk factors with a radically data-driven approach is an impediment to clinical and scientific progress and will fail to deliver the desired results. Critically, this is not an argument against machine learning. To appreciate how machine learning can accelerate and supplement theory-building requires that we dispel the common misconception that machine learning will always produce "black box" models that reveal nothing about how and why they work (cf. Bennett, Silverstein, & Niv, 2019). This review of machine learning fundamentals and common applications is intended to highlight places of potential synthesis between theory- and data-driven approaches to suicidology—and predictive analytics in general—while clarifying what machine learning is and presenting strategies for balancing the opacity of the box with predictive performance.

Machine learning is a growing interest in many areas of clinical psychological science, with the term "machine learning" appearing in the titles of more than 1350 peer-reviewed articles in psychology (per a PsychInfo search conducted August 2020). In addition to suicidology, various machine learning approaches have been used in prediction of neurodegenerative diseases (Moradi et al., 2015), psychosis (Cohen

* Corresponding author.
E-mail address: chriscox@lsu.edu (C.R. Cox).

https://doi.org/10.1016/j.cpr.2020.101940
Received 23 May 2020; Received in revised form 1 September 2020; Accepted 20 October 2020
Available online 23 October 2020

et al., in press), autism (Thabtah, 2019), depression (Kessler et al., 2016), as well as pharmacological (Chekroud et al., 2016), and psychosocial treatment response (Delgadillo & Gonzalez Salas Duhne, 2020). While machine learning is a topic of broad interest, suicide prediction is the first application to which it has been applied at scale in practice, for example in the Veterans Health Administration's Recovery Engagement and Coordination for Health–Veterans Enhanced Treatment (REACH VET) program to help predict suicide and other patient health concerns (Nock et al., 2018; VA, 2017). It is no wonder the deployment of new tools for suicide prediction has been expedited: over 800,000 people worldwide die by suicide each year (World Health Organization, 2017). Approximately 47,000 of these cases are in the United States (Center for Disease Control, 2018), where the rate of suicide is on an upward trend (Hedegaard, Curtin, & Warner, 2018) despite numerous surveillance and prevention efforts (Franklin et al., 2017).

Though the sense of urgency is clear, suicide prediction also distinguishes itself by posing uniquely severe challenges relative to other areas of psychopathology. For one, both the failure to identify risk (false negative) and the misidentification of high or imminent risk (false positive) have serious consequences. Failing to intervene when risk is high can result in death or serious injury, but unnecessary intervention, which may involve hospitalization, can be iatrogenic (see Ward-Ciesielski & Rizvi, 2020). Conventional self-report measures tend to underreport risk and lead to false negatives (e.g., Busch, Fawcett, & Jacobs, 2003), perhaps because at-risk individuals conceal their suicidal ideation for fear of hospitalization. Alternatively, it may reflect the temporal and time-limited nature of suicidal crises and the relative infrequency of clinical assessment (Kleiman et al., 2017). Machine learning has the potential to update risk models based on data that can be acquired much more frequently and which do not rely heavily on self-report, improving the sensitivity of risk assessment and the ability to detect suicidal crises.

Suicide is devastating to those it touches, and tragically ranks among the top ten causes of death in the US (CDC). Nevertheless, even among psychiatric inpatients, suicide is an extremely rare event. Thus, the probability of any individual escalating to a suicide attempt is very low. Risk assessment that is both sensitive and specific in its predictions will require a wealth of evidence—a very strong indicator, or an aggregation of a great many weaker ones. Identifying such indicators is the aim of machine learning but can be challenging. From the perspective of a model that aims to maximize accuracy under uncertainty based on a large number of variables that may or may not carry generally useful information, the "safe bet" is to err on the side of low risk—predict that the current case will not self-harm. Effective suicide prediction must bear all these challenges. Fortunately, there are strategies for confronting each challenge—even if they are not always surmountable. Therefore, by focusing on applications in suicidology, we hope to be directing our efforts where the need is greatest and contribute to the development of methods and rigorous protocols that can be adopted more broadly.

Especially in light of the challenges facing suicide prediction, it is important to clarify our terms. Even though the models being fit are often binary classifiers and the apparent goal is to identify individuals in crisis, models will tend to output a score that corresponds to the probability of belonging to a category (e.g., low risk for suicide attempt, high risk for suicide attempt). While the decision criterion for a binary classifier is often taken to be 0.5, it does not need to be. The threshold is varied across the whole range when computing the *area under the curve* (AUC), which refers to the *receiver operating characteristic* (ROC) curve created by plotting the rates of true positives and false positives against each other at different probability thresholds. A case assigned a high probability is, according to the model, at higher risk of self-injury. This *risk* does not become a discrete *prediction* until a decision criterion is applied.

We will begin with a tear-down of machine learning essentials, followed by an intuitive overview of how various machine learning technologies work—in most cases, what goes by "machine learning" is simply an extension of linear regression. We will then discuss the interface between machine learning and scientific knowledge; depending on the objectives of the project, machine learning can produce models that are interpretable and generalizable, and there are ways to integrate prior expectations and theory within machine learning efforts. We will do this by discussing multiple approaches to machine learning, including those that assume linearity and those that do not. We will conclude by considering the unique challenges facing predictive modeling of low prevalence events, such as suicide, keeping in mind the myriad of objectives and machine learning approaches discussed previously. We also will discuss how future applications of machine learning may confront these challenges to answer important research questions, aid suicide risk assessment, and improve broad suicide prevention techniques.

2. Recipe for a "Machine Learner"

Let us first address what it means for a machine to learn. The language implies a sort of intentional, creative intelligence. This is not so. Machine learning is simply a computer following a recipe, and all such recipes boil down to two essential elements:

1. An **objective**, which is defined in terms of:
   - A definition of how **prediction error** is quantified. Reducing error according to this definition will move the model toward the objective. Prediction error describes the discrepancy between the model's expectation given the data it has provided and the ground truth. While the model is "learning", these discrepancies will be shrinking—at least with respect to the cases it is trained on.
   - An optional **penalty term** to apply downward pressure on model degrees of freedom. Rather than being applied to the model predictions, the penalty is applied to the model itself. By analogy, consider adjusted R^2 or the Bayesian Information Criterion (BIC) when evaluating regression models. Both include a penalty that scales with the number of regressor variables, such that if two models explain a similar proportion of variance in the dependent variable, the one with fewer parameters is preferred.
2. A **compass**, which is defined in terms of an **optimization routine.** Machine learning involves estimating where the objective is and moving toward it.

In other words, a "machine learner" is a computer program with an objective and a specific way to obtain or discover that objective.

Linear regression can be construed as a simple example of machine learning. Variance in a dependent variable is modeled as the weighted sum of one or more independent variables. When we refer to a model, we mean the set of weights applied to the independent variables. Independent variable values are multiplied with their weight, and the sum is the model's prediction about the dependent variable. An ordinary least squares (OLS) regression model minimizes the sum of squared residuals, where a residual is the difference between the true and model-predicted value of the dependent variable.

Regression is a statistical procedure that solves this problem directly, by applying specific equations consistent with probability theory. Thus, a regression model supports statistical inference about each of the independent variables. The same model can also be obtained with a machine learning approach. The model would be initialized with random weights, and thus it would begin by making random predictions. These predictions would correspond poorly to the known values of the dependent variables. In machine learning there are many ways this error might be quantified, and they will lead to the discovery of different models. To discover the same model regression would produce, the error should be quantified as the sum of squared differences between the true and model-predicted values. Thus, the machine learning **objective** would be to adjust the model incrementally to discover weights that

minimize error defined in this way. To achieve this objective, the model requires a **compass**, which would be an optimization routine such as gradient descent. So, instead of obtaining the weights in one step by applying statistically-motivated equations as is done with regression, machine learning will search for a model that achieves an objective (such as minimizing squared-residual differences) by following a compass (such as gradient descent). This process of observing data, making predictions, evaluating error, and updating the model, is called *training* in machine learning.

Linear regression has been used as a statistical model to identify factors that explain statistically significant amounts of unique variance in suicide behavior (e.g., Joiner Jr et al., 2005). However, decades of research has produced hundreds of investigations with statistically significant results with few or none reaching a level of significance that impacts clinical practice (Franklin et al., 2017). Following the preceding description of machine learning, used to approximate what regression already provides, it may not be clear what advantage machine learning has over regression. What is the field doing now with machine learning that differs from prior work with regression? The key is that machine learning can discover models that regression cannot. Regression is restricted because it is designed to support statistical inference; models that support statistical inference correspond to a small subset of all possible models that may exclude models that maximize prediction performance. Machine learning searches this larger set of models to discover ones that maximize performance. By removing the limits demanded by statistical theory, machine learning can develop predictive models based on datasets far larger and more complex than would be possible with linear regression. The compromise is that it is more difficult to infer the general significance of individual independent variables within models discovered through machine learning.

3. Supervised vs. unsupervised machine learning

Most cases of predictive modeling involve *supervised* machine learning. To unpack the metaphor, a machine learner is "supervised" if the values it produces during training are checked against some target values, and the learner is provided with feedback. The dependent variable can be thought of as a teacher supervising the machine learner as it assigns labels or numeric outputs to the inputs it receives. When the learner makes a mistake, the teacher points it out and provides instruction about what the label should have been. In the context of suicide research, the "supervisor" may be a dependent variable that codes whether an individual has attempted suicide. The majority of machine learning approaches in suicidology have used a supervised approach, whether predicting suicide attempts of suicide death (e.g., Kessler et al., 2017; Walsh, Ribeiro, & Franklin, 2017).

Unsupervised machine learning, on the other hand, is not guided by such feedback. Instead, unsupervised learners organize the data without explicit guidance. This can be quite revealing, especially in datasets with many variables associated with each case. For example, principal components analysis, independent components analysis, and factor analysis are unsupervised techniques for summarizing variance that is shared by multiple variables. They all project a large space of variables onto a smaller set of "basis functions" or "latent variables", which can reveal interesting structure and clarify the relationship between the observed data and theoretical constructs of interest. They differ with respect to what the "goal" of the optimization is, and so will summarize the same data in different ways. For example, there may be several variables that correlate with the presence of suicidal thoughts, while others correlate with having a suicide plan. The presence of suicidal thoughts and the presence of a suicide plan are latent variables that are difficult to assess directly, but can be estimated with different standard assessments or, potentially, revealed in the correlations among many variables in large medical datasets by unsupervised machine learning. The researcher performing a principle components analysis, for example, is responsible for identifying if any of the principle components identify in the data correspond with latent variables of theoretical or practical interest.

On the other hand, *clustering* algorithms can reveal structure in the similarity among cases or variables. Especially when cases are defined in terms of many variables, it can be difficult to appreciate how cases relate to one another when inspecting the raw data. There are many clustering algorithms, and all behave differently, but for illustrative purposes we will consider two that are widely used: *k*-means and hierarchical agglomerative clustering. Fig. 1 serves as a visual aid. *K*-means clustering will divide cases into *k* mutually exclusive sets, where each set is defined with reference to an anchor point. Agglomerative hierarchical clustering is a related unsupervised technique that does not return a specific number of groups and does not require an initial estimate about how many groups may exist. The results of hierarchical clustering are typically displayed in a dendrogram, where the height of the connection between groups corresponds to the distance between them. The dendrogram can be cut at any height to discretize the data into clusters, where higher cuts will result in fewer clusters that are more distant from each other. Both methods are simple and unsupervised but note that both involve human intervention: for *k*-means, *k* must be chosen, and a hierarchical clustering analysis will not yield clusters until a decision has been made about where to cut the dendrogram. Clustering may reveal sets of cases distributed through a database with consistent similarities among measured variables. Likewise, it may reveal variables that tend to be correlated across cases. While a relatively descriptive analysis on its own, clustering can be a useful tool for facilitating research within large datasets when paired with other methods.

4. Supervised machine learning for generality and insight with linear models

While we will emphasize alternatives to "black box" machine learning, this stereotype is not baseless. When a machine learner is given free-reign to build a predictive model from any and all combinations of available data, and the available data are not pre-selected by hypothesis, the resulting model will often defy principled interpretation and thus may be less useful in clinical science investigations regarding why people die by suicide. This is not without its merits. A limited-application black-box model that is demonstrably sensitive and specific will be a major achievement for suicidology, even if it cannot be ported to other hospitals/clinics directly to aid suicide prediction. Scientific progress, however, requires replicability and understanding of underlying mechanisms. Fortunately, the machine learning recipe permits multiple simultaneous goals, in addition to predictive accuracy.

Consider a predictive model implemented at a large medical center that improves detection of individuals who will attempt suicide within two-weeks of discharge. The objective is to enhance risk assessment, not identify malleable treatment targets to prevent suicide. Given the objective, whatever machine learning approach best supports risk assessment should be used, regardless of whether the model lends itself to interpretation that may enhance the scientific understanding of why a patient sees suicide as an option, or what progresses suicidal thoughts to actions (Klonsky, Saffer, & Bryan, 2018). On the other hand, by setting additional objectives, it may be possible to apply the model or insights gleaned from the model to applications at other medical centers, community mental health offices, etc., or to inform theories of the etiology and maintenance of STBs.

In what follows, we will discuss several "additional objectives" and how they contribute to making more generalizable and interpretable models. The methods we will discuss are summarized in Fig. 2, which depicts nine variables belonging to three mutually exclusive groups based on prior expectations about the dataset. For example, there may be distinct groups of variables that indicate access to lethal means, relate to predictors of suicidal ideation (but likely unrelated to lethal means access), and refer to immunization history (or any other set of variables that cluster within the dataset, but which do not predict suicide). The top row (depicting the "ground truth") indicates that variables in the red

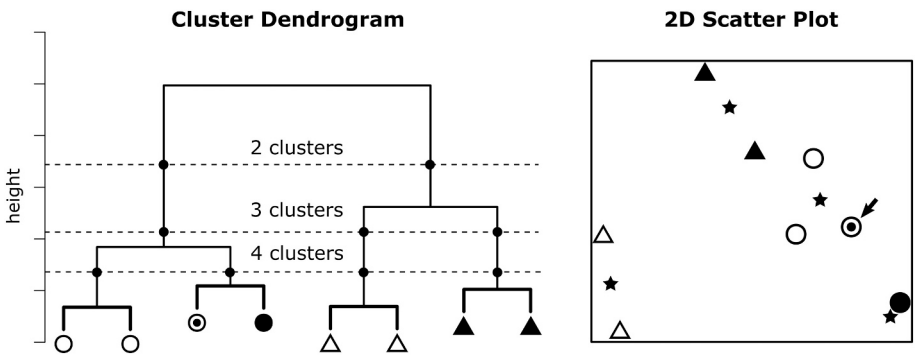

Fig. 1. Unsupervised hierarchical and k-means clustering.
Note. Fill and shape distinguish items as they cluster into 4 categories. The bulls-eye symbol indicates an item that k-means clusters with the other open circles, but the hierarchical cluster analysis groups with the other filled circle. See the main text for an explanation of each method. Based on the "height" at which the dendrogram is "cut", different numbers of groups are implied (dotted lines on histogram). K-means clustering involves placing k "centers" (represented as stars in the 2D scatter plot) and assigning points to clusters based on which center (i.e., star) they are closest to. We depict a k-means solution corresponding to k = 4. Each method can lead to different solutions (as identified by the arrow in the 2D plot).

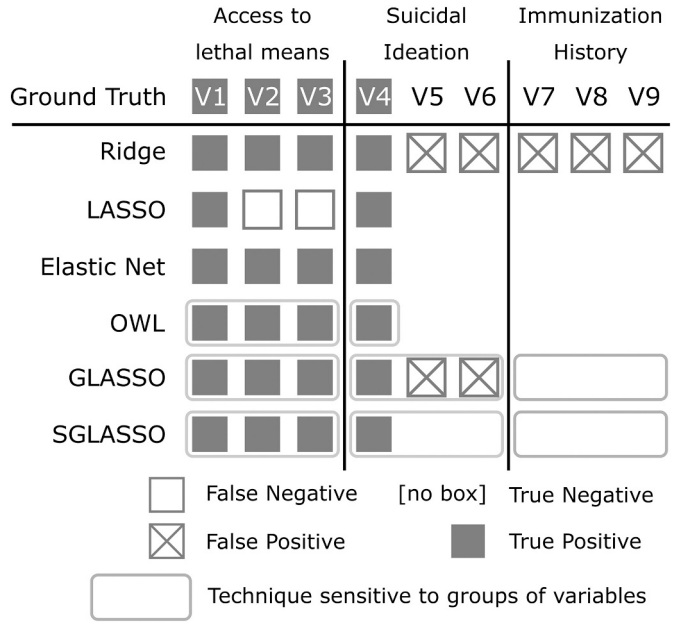

Fig. 2. Variable selection and grouping with regularized linear machine learning.
Note. Nine variables from 3 groups are depicted as columns. The first row indicates which variables contain useful information in this dataset (Ground Truth). Ridge regression does not discriminate between informative and uninformative variables, resulting in many false positives. LASSO enforces sparsity aggressively, leading to two false negatives. By mixing Ridge and LASSO, Elastic Net may identify all informative variables and exclude uninformative ones but does not signal that the 3 "lethal means" variables form a group. Ordered-weighted LASSO (OWL) behaves like Elastic Net but is designed to signal variable groupings more clearly. Group (G)LASSO and Sparse Group (SG) LASSO allow the researcher to specify these groups a priori rather than inferring them from the data.

group all contribute useful information, but only one of the variables in the green group does and none of the variables in the blue group do. In subsequent rows, units are shaded if the associated machine learning method will assign that variable a non-zero coefficient (i.e., suggest that the variable is important for prediction) under ideal circumstances. Variables are shaded and colored when the machine learning method also conveys or includes information that allows the underlying variable groups to be distinguished. The purpose of the figure is not to advocate for one machine learning approach over another, but to support basic intuitions about interpretations supported by each.

4.1. A note on degrees of freedom and overfitting

The goal of machine learning is to discover a model that can make accurate predictions given new data. A model should assign weight to variables that reliably relate to the outcome of interest and which form a sturdy basis for generating accurate predictions. Critically, the goal is not to minimize prediction error on the *training set*—the set of cases for which the dependent variable has been measured and is being used to "supervise" the machine learning. Every variable added to the model potentially adds an entirely new *degree of freedom*. A model's flexibility increases with its degrees of freedom. The more flexible a model is, the better it will fit its training set. However, that flexibility allows the model to conform to subtle idiosyncrasies of the particular sample that constitutes the training set that are not representative of the population it was drawn from. In other words, the model will be informed by variables that, with respect to the population, are pure noise. This is called *overfitting* and it occurs when spurious probabilistic relationships between a dependent variable (e.g., suicide attempt) and the variables being modeled (e.g., electronic health records) exist in the training set. The risk of this increases as the ratio of variables approaches and exceeds the number of cases. Counterintuitively, models that perform extremely well on the training set often generalize poorly compared to models that perform moderately on the training set. The latter model has managed to ignore the noise while identifying variables the are genuinely important in the population.

As alluded to in our recipe for a machine learner, the qualities of a model obtained through machine learning will be determined by the objectives sought during optimization. By expanding the recipe, we can introduce competing objectives and expand the definition of what constitutes a "good model". For instance, to avoid overfitting to the training set, a penalty can be applied that forces the model to use only some of the degrees of freedom available to it.

4.2. Applying pressure to constrain model degrees of freedom

A common way to constrain model degrees of freedom and thereby combat overfitting is through *regularization*, which is a general term referring to a variety of penalties applied to the model weights. Reducing this penalty is a secondary objective, sought alongside the minimization of prediction error. *Regularized regression* techniques avoid overfitting by introducing a penalty such that weights cannot be set arbitrarily to maximize predictive accuracy, which forces the machine learner to construct a simpler model that emphasizes the most important aspects of the dataset. Thus, the competing objectives are to 1) maximize accuracy and 2) utilize the fewest degrees of freedom. Exactly how this secondary objective is implemented will lead to qualitatively different models. We will begin by contrasting two such penalties: the sum of squared weights (ridge regression; Hoerl & Kennard, 1970) and the sum of the absolute values of the weights (LASSO; Tibshirani, 1996).

Penalizing the sum of squared weights will produce models where every variable is assigned a small (but nonzero) weight. This reason for this is intuitive: by squaring weights, large ones are penalized heavily, but weights approaching zero have insignificantly small penalties. If a variable is very predictive, rather than placing a large weight on that one variable, the weight is spread out over less predictive but correlated variables in the data. The reason for this is also intuitive: $2^2 + 0^2 = 4$, but $1^2 + 1^2 = 2$. These properties of the penalty allow ridge regression to produce useful models to problems where there are large numbers of variables that have complex covariance structure (Marquardt & Snee, 1975), even when there are more variables than cases to classify, which means it can be applied in cases where standard regression cannot. Despite these useful qualities, ridge regression produces "black box" models that are difficult to interpret because they do not differentiate between informative and uninformative variables (all variables receive nonzero weights) and the size of the weight assigned to a variable by ridge regression does not transparently reflect that variable's importance. A variable may be associated with an exceptionally small coefficient because it contributes very little to model prediction, or because it is important but shares variance with many other variables. Thus, ridge regression is a "black box" solution for constraining model degrees of freedom, which promotes generalizability but not interpretability.

Conversely, penalizing the sum of the absolute weights will produce *sparse* models with many weights set to zero. Like ridge regression, LASSO constrains model degrees of freedom and trades accuracy on the training set for a simpler solution that is less likely to overfit the training set and thus generalize better to related cases. Unlike ridge regression, the weights assigned by LASSO have clear interpretations: a zero means the variable does not contribute to prediction; a larger weight means the variable contributes more to prediction. Where ridge regression was pressured to spread weight among correlated variables, LASSO does the opposite. Given a set of correlated variables, the one that predicts the dependent variable (e.g., SITB risk) best gets *all* the weight, and the others should be set to zero. When regularizing with LASSO, there is no advantage to sharing weight between correlated units: $2 + 0 = 2$ and $1 + 1 = 2$. However, because one of those two variables in the example is going to be a slightly better predictor of STBs in the training set, spreading weight to the lesser variable will incur the same penalty on the weights for lower accuracy. Thus, LASSO will tend to select only the most strongly predictive variables. In terms of variable selection, LASSO tends to be specific, but not sensitive. It identifies a small set of variables that tend to be a subset of those that really do convey useful predictive information, and largely avoids uninformative variables.

Not all problems call for the same intensity of regularization. In some datasets, nearly all variables may convey some degree of independently relevant information, while in others most variables are irrelevant to prediction. The intensity of regularization, and the relative importance of regularization penalties when multiple are applied to the same model, can be dialed up or down via *hyperparameters*. Hyperparameters are values that influence how a model is fit but is not assigned to any particular variable nor used directly to generate predictions once the model has been fit. By selecting appropriate hyperparameters, the effectiveness of regularize regression can be greatly enhanced.

By including a competing objective that constrains model degrees of freedom, machine learning can construct models from large sets of variables, many of which are expected to be uninformative. This has enabled innovative work attempting to predict suicide attempts based on social media data, exploring unconventional factors such as total number of posts, frequency of posts, latency between posts, syntax and sentence construction choices, and many others that may or may not correlate with STBs (Cheng, Li, Kwok, Zhu, & Yip, 2017).

4.3. Applying pressure to identify structure among variables

More complicated regularization is also possible. Elastic net, for example, includes both the LASSO and ridge penalties and a *hyperparameter* that scales their relative importance. In other words, LASSO and ridge regression can be implemented as special cases of elastic net by setting this hyperparameter such that either the LASSO or ridge penalty is multiplied by zero. When it works well, elastic net balances performance and interpretability better than either LASSO or ridge regression can independently by encouraging solutions with *structured sparsity*. While LASSO encourages sparse solutions at the level of individual variables, penalties that encourage structured sparsity encourage the selection of sets of related variables. The aim of elastic net is to set many variables to zero while assigning weight to a small number of variable sets, where the variables in each set correlate with each other and make similar contributions to model performance. This grouping of similar variables is desirable from both practical and scientific perspectives. Practically, it adds some redundancy to the model variable space, which pure LASSO would have excised in the single-minded pursuit of sparsity. This can improve the model's generalizability, if the most influential variable in a variable set—the one LASSO is biased to select while shrinking the coefficients of the other correlated variables to zero—is not the most influential variable in another dataset where the model will be applied. Scientifically, identifying a set of related yet predictive variables provides a more solid basis for hypothesizing about the reason why the variables are predictive than considering a single variable out of context of from the set. For example, the risk of suicidal thoughts and risk for suicidal behaviors are associated with different sets of factors (Klonsky et al., 2018). Factors within each set likely explain overlapping variance in patient risk, in addition to a small amount of unique variance. When taken together, these factors are associated with only a small subset of datapoints that are likely to be available for each patient within a clinical database. Ridge regression would allocate some weight to all variables in the database; LASSO would allocate weight to only a fraction of the variables that are truly associated with thoughts and behavior. Elastic net is designed to discover a more "complete" model: in this example, it would distribute weight over the correlated variables within the ideation and risk variable sets while still excluding irrelevant variables from the model.

Elastic net (Kessler et al., 2017) and pure LASSO (G. E. Simon et al., 2018; Tran et al., 2014) have been applied in the suicide prediction literature, but to the best of our knowledge other methods that promote structured sparsity are yet to be explored. For example, ordered weighted LASSO (OWL; Figueiredo and Nowak (2016)) is similar to elastic net but puts additional emphasis on clearly delineating the sets of variables that it discovers. Unlike an unsupervised principal component or clustering analysis, only clusters of *predictive* variables will be returned by OWL methods.

4.4. Applying pressure to incorporate prior knowledge

Elastic net and OWL obtain solutions with structured sparsity based on the covariance among variables that the model is trained on. Alternatively, there are regularization techniques that allow the researcher to incorporate prior expectations about the relationships among variables into linear regression, rather than relying on the model to identify them. Variables assigned to the same group will be considered similar. To the extent that group assignments reflect the true structure of the data, this will lead to improved model fit while selecting interpretable sets of variables. Defining groups that are inconsistent with the structure of the data will lead to inferior model fits. Thus, grouping variables that are believed to underly a similar construct, such as any of the many that are proposed by the Integrated Motivational Volitional theory of suicide (O'Connor, Cleare, Eschle, Wetherall, & Kirtley, 2016), the interpersonal theory of suicide (Van Orden et al., 2010), or the three-step theory (Klonsky & May, 2015), is a way of implementing hypotheses within the machine learning framework.

There are several regularization penalties that accommodate prior expectations in this way. Group LASSO (Yuan & Lin, 2006) stipulates that including one variable from a group requires that all fellow group

members receive non-zero weight as well. Sparse group LASSO (N. Simon, Friedman, Hastie, & Tibshirani, 2013) and sparse overlapping groups LASSO (Rao, Cox, Nowak, & Rogers, 2013; Rao, Nowak, Cox, & Rogers, 2016) regularize over groups and individual variables to encourage more flexible patterns of structured sparsity that do not have this all-or-nothing quality with respect to groups. Models that permit overlapping groups allow a single variable to relate to multiple constructs.

4.5. Garbage in, garbage out

Machine learning excels at discovering patterns and relationships in large datasets, but even in the most data-driven research there is an essential need for curation and common sense to exclude obviously problematic variables that may undermine a model's utility. To take an extreme example, imagine if the medical records a model is trained on include a cause of death report. This variable may perfectly predict whether a case died by suicide, but clearly this model is useless for the intended purpose of evaluating the risk of self-harm with enough time to intervene.

4.6. Alternative solutions to the same problems

There are three notable omissions from this discussion: linear discriminant analysis (LDA), support vector machines (SVMs), and Bayesian models. LDA is similar enough to logistic regression that it is subsumed by that discussion; in practice logistic regression is likely preferable (Liong & Foo, 2013). Support vector machines (SVMs) are a classification technique that can be either linear or nonlinear. In the linear case, SVM is like LASSO-regularized logistic regression, in that a sparse set of variables will be identified as informative, but SVMs seek a different objective. Because LASSO and SVMs are functionally equivalent (Jaggi, 2014) we will allow our discussion of LASSO to also serve as a discussion of the linear SVM, even though they obtain their solutions through conceptually and computationally different optimizations.

The supervised machine learning techniques surveyed above can all be implemented in a Bayesian framework. For example, from a Bayesian perspective, regularization incorporates prior knowledge or expectations about the distributions from which model parameters are sampled. The prior knowledge being incorporated with ridge regression is that many variables share variance with a relatively small number of informative latent dimensions, and the knowledge being incorporated with LASSO is that only a few variables are relevant. Bayesian (ridge) regression achieves this via the prior belief that parameters arise from a normal distribution of some width centered on zero—more narrow distributions imply more severe regularization (i.e., a greater expectation that the true parameter value is near zero). Bayesian LASSO is similar, except with the normal distribution replaced by a Laplacian distribution. Hierarchical Bayesian models can be designed to perform like group regularized regression and thus achieve structured sparsity (Chen, Chu, Yuan, & Wu, 2016; Nathoo, Greenlaw, & Lesperance, 2016).

Despite these parallels, Bayesian machine learning is a distinct approach to avoiding overfitting with advantages and disadvantages relative to "conventional" machine learning, not merely an alternative approach to regularization. For example, Bayesian priors allow virtually any prior belief to be expressed (not just that values should be small/zero). This *increases* researcher flexibility while *decreasing* model flexibility: the more specific and accurate the prior beliefs are, the less the risk of overfitting. The cost for this flexibility is that determining appropriate priors for complex models is daunting and permits virtually infinite experimenter degrees of freedom, while fitting Bayesian models tends to be more complex involve more complex optimizations. For the purposes of this review, and because it is currently uncommon in the suicide prediction literature (Belsher et al., 2019), we will not discuss Bayesian machine learning any further.

5. Nonlinear models of suicide

To this point, we have focused on linear models. When it comes to interpretability and generalization, linear models have a distinct advantage over nonlinear models in that they are defined explicitly in terms of the input variables. Each input variable is associated with a single weight, the same weights always apply to all cases, and the model prediction is simply a weighted sum of those input variables. The combination of information from different variables is linear. A nonlinear model has more flexibility when combining information across variables and may develop abstractions or involve a series of discrete decisions. For instance, neural networks and nonlinear SVMs re-represent the variables provided in terms of complex combinations that can obscure the relationship between the solution and the original variables, while decision trees are composed of a series of classifiers and apply different weights to different subsets of the data. Consequently, the risk of overfitting the training set is often greater with nonlinear models. In the following, we will briefly discuss the nonlinear machine learning techniques that have been applied to suicide prediction.

5.1. Neural networks

Neural networks comprise a broad and flexible class of nonlinear models and, consequently, have the potential to be extremely complex. The exploding field of "deep learning" is driven by neural network models with multiple processing layers, where each successive layer further abstracts from the variables driving prediction. They have the advantage of being incredibly powerful: neural networks are universal approximators (Hornik, 1991; Hornik, Stinchcombe, & White, 1989), which means that in principle they can learn to implement any function necessary to relate between the inputs and target outputs they are trained on. This comes with a major caveat: a neural network does not tell you what function it has implemented, and the data transformations may be extremely complex and opaque to the researcher. When machine learning is accused of producing black box models, neural networks are the primary culprit. Neural networks are rarely preferable to other methods when the goal is to develop general insights. In the rare cases that neural networks have been applied to suicidology, they have made less accurate predictions than simpler, linear methods (e.g., Amini, Ahmadinia, Poorolajal, & Moqaddasi Amiri, 2016).

5.2. Decision trees

Decision trees, and the associated random forests, are nonlinear in a different way. When constructing a decision tree model, one begins with the full dataset and then picks a variable that allows the data to be split. This split will, as best as possible, distinguish between those at high and low risk for suicide—but more than likely this first split based on one variable is not going to provide a satisfactory solution (Hill, Oosterhoff, & Kaplow, 2017). Therefore, the process is repeated within each partition of the dataset after the first split, and so on until a satisfactory solution is obtained. The nonlinearity of the approach comes from the fact that the variable chosen to perform each subsequent split is based on a different subset of cases. A decision tree allows, for example, different variables to be predictive of suicide for individuals making more or less than $100 k a year. In a linear regression approach, a single set of weights are obtained that make predictions for all items. In a decision tree analysis, a different set of variables leads to each terminal "leaf" of the tree.

On one hand, decision trees lend themselves to clinical decision making, which is frequently binary in nature. Each bifurcation of the tree is associated with a decision point dictated by a single variable, and by following a path of simple decisions it is possible to visualize the inner workings of the model and sort cases into high and low risk (or any other categorization of interest). However, at each step in the tree, fewer and fewer cases are being sorted out based on effectively the same large

variable space. Thus, each subsequent split has an increased risk of being overfit to the data. A common solution to the known issue of overfitting by decision trees is to produce "random forests", which involves running many decision trees on the same dataset, where each tree is built with respect to a different subset of the cases or variables. The final decision is taken as the "majority vote" of the individual decision trees. While the random forest approach can guard against overfitting and provide good predictions that are more likely to generalize to another dataset, it eliminates the simplicity and transparency of the decision process. Therefore, it is difficult to interpret how or why random forests work, relative to decision trees or linear models.

6. Summary

Nonlinear machine learning technology like random forests and neural networks have been applied to build suicide prediction models. While the motivation to use such technology is intuitively appealing ("there might be complex structure in the data, so we ought to use a sophisticated model in order to detect it!"), in practice nonlinear models do not tend to significantly outperform linear models. This can be seen with respect to the area under the curve values reported in Beshler et al.'s (2019) review and is also true in other fields. In cognitive neuroscience, for example, a push to apply machine learning to patterns of functional brain activity resulted in nonlinear models being applied early on (Cox & Savoy, 2003; Davatzikos et al., 2005; Hanson, Matsuka, & Haxby, 2004), but quickly a consensus formed that the added complexity and loss of interpretability was not worth the negligible or non-existent performance gains (Kamitani & Tong, 2005; Pereira, Mitchell, & Botvinick, 2009). However, the objective of machine learning in cognitive neuroscience is to evaluate hypotheses, not to create brain computer interfaces or any other practical tool where achieving the best possible readout might be preferred over interpretability. Our recommendation of linear over nonlinear models is consistent with other authors and grounded in our hope that machine learning will be used deliberately and strategically to advance theory and understanding of suicide in general—especially because insights from modeling available data may inform the development of superior datasets through enhanced assessment and data collection practices.

7. Machine learning and suicidology

Machine learning can serve a useful scientific role in a feedback loop of hypothesis, model building, and theory development. The preceding review of machine learning was presented to establish this thesis by clarifying the approach and cultivating several key intuitions:

1. Machine learning refers to a wide range of data driven techniques that are sensitive to probabilistic regularities in large datasets with many potential predictors. Therefore, it is important to be mindful that "when all you have is a hammer, everything looks like a nail" and seek to establish a breadth of knowledge about machine learning technology before launching a project that relies on machine learning.
2. Machine learning methods can emphasize predictive accuracy at the potential cost of interpretability/generalizability, or vice versa. Therefore, decisions about what methods to use should appear at all stages of project design, especially in the early stages when the goals of the project are being discussed.
3. Machine learning can be applied with a variety of objectives in mind, including constructing models of complex datasets that are useful for developing and testing hypotheses in suicidology. Therefore, important progress can be made by considering how relevant hypotheses and domain expertise can be translated into the design and calibration of machine learning tools.

These intuitions, coupled with our comments on specific machine learning techniques, may serve as a foundation for making the approach more accessible and promote its application in a range of contexts. We now turn a discussion to some of the challenges involved in modeling STBs. This discussion will be organized around two major themes: input selection and the low prevalence of suicide in electronic health records (EHR) and within the general population.

7.1. Input selection

A suicide prediction model is only as good as the data it is trained on, and acquiring good data relies on knowing the factors that motivate and escalate STBs. In a recent review of suicide risk factors presented in the literature over the past 50 years, Franklin et al. (2017) revealed that many models might be constructed based on current the current state of theory, and none of them support suicide prediction in practice. Machine learning represents a possible way forward, as it enables the exploration of large variable spaces, including demographic information, clinical diagnoses, healthcare utilization records, medications, socioeconomic status, physical health and vital signs, prior suicide attempts, and more. Exploratory analyses of this kind have begun to emerge in the literature (McCarthy et al., 2015), including work that considers the contributions of predictive factors at multiple time-scales relative to suicide attempts (Walsh et al., 2017). However, the models constructed in these studies were based on enormous variables sets, and machine learning methods were employed that emphasized prediction over the interpretation of individual variables—although Walsh et al. (2017) do report interesting changes in variable importance as a function of how far in the future a suicide attempt is. Other work has employed relatively smaller and more controlled variable sets, which allows clearer conclusions to be drawn about their importance to suicide prediction. For instance, survey data that assesses suicide risk factors improves suicide prediction relative to models trained only on EHR administrative data (Bernecker et al., 2018). EHR data has the benefit of using readily available data but may lack critical information which puts an upper limit on the predictive ability of models trained only on this information, but these limits can be pushed upward with relatively simple, theoretically motivated data contributions. Thus, the most productive way forward may be an iterative approach that involves a combination of exploratory and hypothetically motivated machine learning efforts.

Techniques like group LASSO that allow the researcher to specify groups of related variables a priori can also help with the input selection problem. On the exploratory side, grouping variables based on clinical expertise and prior research can prevent overfitting, reduce the variance in variable weights over model fits, and guard against drawing interpretations based on overly sparse variable selections. In hypothesis driven models, applying regularization in this way allows one to translate important aspects of the hypothesis into the machine learning framework, and allows one to test whether relating groups of variables to constructs in particular ways leads to superior model fits. This might be applied to evaluate the construct validity of theoretical models within the ideation-to-action framework (Klonsky, May, and Saffer, 2016): the number of potential risk factors is large yet structured by hypothesis. Machine learning is being actively applied to this question, and exciting progress is being made exploring the space of risk factors (Klonsky et al., 2018; Nock et al., 2018). Folding hypothesis and prior expectations directly into the learning progress may aid future work in this area.

A major challenge for suicide prediction is that suicide intent can vary dramatically over the course of hours and days (Kleiman et al., 2017; Torous et al., 2018). While EHRs and targeted surveys and interviews may reveal information about elevated risk, the events that trigger the escalation from thought to action may operate at far finer temporal scale. A transformative goal of machine learning within suicidology is to leverage novel and unconventional sources of data that operate at a scale that may be more appropriate for detecting the escalation. Publicly available natural language datasets, like those

generated organically via social media platforms like Twitter, may contain a wealth of information that correlates with STBs. There is some evidence that expressing thoughts of suicide on social media correlates with clinically assessed anxiety scores (AUC = 0.75) and suicidal probability (AUC = 0.61; Cheng et al. (2017)). On one hand, this a useful proof of concept that self-expression on social media corresponds to established clinical risk factors, and on the other may suggest a "social media assessment" for suicide risk. Others have begun work categorizing the ways that suicide is referenced on social media, including seeking information or support to aid in developing a plan for suicide, expressions of suicidal intent, memorial references to life events, and flippant references to suicide, to lay the foundations for further natural language processing of expressions of suicidal ideation on social media (Burnap, Colombo, Amery, Hodorog, & Scourfield, 2017). Discovering informative metrics with higher temporal resolution is an exciting and potentially transformative avenue of future machine learning work in clinical psychology (Cohen et al., 2019) .

Other novel risk factors may lie in patterns of neural activity elicited by the evaluation of diagnostic stimuli. Just et al. (2017) was able to accurately identify 94% of individuals who had made a suicide attempt from healthy controls using a machine learning algorithm applied to functional magnetic resonance imaging (fMRI). In this study, fMRI was used to determine the neural signatures of concepts related to death and life, and found that "death, cruelty, trouble, carefree, good, and praise" were most useful concepts in determining those who had made a suicide attempt (Just et al., 2017). While data of this kind is expensive to collect, it has the potential advantage of bypassing self-report and providing more direct access to the patient's internal state.

7.2. The base rate of STBs and its impact on prediction accuracy

Recently, Belsher et al. (2019) wrote that "suicide prediction models produce accurate overall classification models, but their accuracy of predicting a future event is near 0" (pp. E1). This may appear to be a contradiction: how can a model be both highly accurate and fail to predict the future? Unpacking this statement is instructive and highlights an important practical challenge facing predictive modeling for suicide risk.

While the rate of suicide in the general population is tragically high, most people will never attempt suicide. In fact, the base rates of suicide mortality and suicide attempts are substantially lower than the base rates of other health outcomes that predictive models are being constructed anticipate (e.g., cardiovascular episodes). If 99% of the cases in a training set have not attempted suicide, then a model can be 99% accurate by predicting that no one in the training set has attempted suicide. Thus, it is generally understood that accuracy is highly confounded with the prevalence of suicide and is not the metric of most interest when describing model performance. Instead, model performance is reported in terms of sensitivity (the proportion of attempts that the model labeled "high risk"), specificity (the proportion of non-attempts that the model labels "low risk"), and area under the ROC curve (AUC; the ratio of sensitivity to specificity over a range of decision criteria). Sensitivity and specificity (and by extension AUC) are not confounded with prevalence in the same way:

$$sensitivity = \frac{\#correctly\ predicted\ positive\ cases}{\#true\ positive\ cases}$$

$$specificity = \frac{\#correctly\ predicted\ negative\ cases}{\#true\ negative\ cases}$$

That is, the sensitivity metric is completely independent of the number of negative cases, and vice versa for specificity. In addition to these metrics, however, it is common to report positive predictive value (PPV; the proportion of cases labeled "high risk" that correspond to attempts—the inverse of sensitivity). PPV is a metric of great practical interest, because it provides insight into the false alarm rate. A predictive model that frequently recommends expensive treatment to patients that do not need it is undesirable. However, the relationship between PPV and prevalence is less appreciated. While PPV is typically computed directly from information about true and false positives:

$$PPV = \frac{\#correctly\ predicted\ positive\ cases}{\#predicted\ positive\ cases}$$

… the equation can be expanded to reveal that, holding sensitivity and specificity constant, PPV will change as a function of prevalence, and that lower prevalence is associated with less PPV:

$$\text{PPV} = \frac{sensitivity \times prevalance}{(sensitivity \times prevalance) + (1 - specificity)(1 - prevalance)}$$

Thus, a model with excellent sensitivity and specificity may nevertheless have low PPV if applied in a context where the prevalence of the entity being predicted is extremely low—such as the case with suicide mortality and suicide attempts. Because extremely low prevalence is the norm for suicidal behavior, models are typically trained on samples with artificially high prevalence, and therefore the statement made by Belsher et al. (2019) is not an inherent contradiction. Instead, it underscores a major applied challenge for predictive modeling of STB.

To make this concrete, suppose a predictive model is constructed based on a sample of 10,000 cases (5000 experiencing suicidal ideation and 5000 who died by suicide; prevalence = 50%), and this model performs nearly perfectly: sensitivity = 99%; specificity = 99%. Based on the equation above, we also know that, *in this sample*, PPV = 99%. The trained model is then shared with a psychiatrist who works in an inpatient unit which sees suicide at the same rate as most inpatient units: 147 per 100,000 (Madsen, Erlangsen, & Nordentoft, 2017). The psychiatrist runs the model through a retrospective dataset of his patients and finds that that model generalizes beautifully: sensitivity and specificity remain near 99%. However, she finds the PPV is near 13%—out of the cases the model predicted would die by suicide, only 13% did. This follows directly from the equations above, and it indicates that if this extremely accurate model was used to screen her patients, 87% of the cases it flags would be false alarms. From the perspective of the patient, improper application of mental health services can lead to adverse outcomes (Chung et al., 2017; Kapur et al., 2013; Paksarian et al., 2014).

This discussion is not new to the prediction of medical concerns (e.g., Glaros & Kline, 1988), and it places an exceptionally high bar on the quality of predictive models before they can be clinically useful for suicide prediction. There are also alternative readings of this example: very few truly suicidal cases are being missed, and even with 87% false positives, this is a much more manageable load for manual inspection than the full patient population. In other words, this model may still be viewed as a practical triumph that conveys significant value to the practice, even despite a high false positive rate. That said, predictive models of suicide are not performing anywhere near as well as in this contrived example, and the PPV "in the wild" would be expected to be very low, as Belsher et al. (2019) point out.

This raises important methodological considerations: should models be fit to training sets that reflect the prevalence of suicide in the population where the predictive model will be applied? Or, should more balanced samples be used to support the machine learning procedure and provide more confidence in the risk factors that are identified in the model? Recall that machine learning discovers a model by optimizing toward an objective such as reducing prediction error. When most outcomes belong to one category, machine learning can achieve low prediction error simply by always predicting the more likely outcome. This model will have low specificity but provides a reasonable default position if evidence provided by the available variables is weak. When models are trained on more balanced samples, error can only be reduced by assigning weight to variables.

Rather than restricting data collection, conditions can be made more balanced through resampling—either under-sampling cases associated

with the more frequent outcome, or over-sampling cases associated with the rare outcome (Burnaev, Erofeev, & Papanov, 2015; Lee, 2014; Torgo, Branco, Ribeiro, & Pfahringer, 2015; Weiss, 2004). A full consideration of strategies for modeling rare events is beyond the scope of this paper. The point is that it is important to train on data that are representative of the variance within conditions, but modeling on data that is representative of the prevalence of conditions may introduce bias into the solution that minimizes prediction error without actually learning how to identify the cases of interest. It is possible to estimate the PPV of a model under more realistic prevalence assumptions without training on data that exhibits that prevalence. In short, better models may be obtained by training on more balanced sets of examples, and this balance can be achieved through resampling and simulation (Burnaev et al., 2015; Lee, 2014).

7.3. An ounce of prevention…

As a low-probability outcome resulting from myriad, potentially idiosyncratic, factors culminating in a window of acute risk, suicide will always be difficult to accurately predict. While the field strives to identify generally predictive risk factors, advancements are being made in the realm of suicide prevention. For example, the Youth Nominated Support Team Intervention (YST) dramatically reduces mortality among adolescents who recently attempted or contemplated suicide without negative side effects by training a trustworthy adult, nominated by the at-risk adolescent, in treatment support strategies (King et al., 2019). Likewise, preparing those at risk for suicide with a list of coping skills and supportive resources and following-up over the phone has been shown to reduce suicide attempts by 45% while doubling the likelihood of seeking outpatient mental health services (Safety Planning; Stanley & Brown, 2012; Stanley et al., 2018). Even indirect initiatives, such as depression and suicide awareness campaigns in hospitals waged with pamphlets, are related to fewer deaths by suicide in the following months (Matsubayashi, Ueda, & Sawada, 2014).

Interventions like these are low-cost (compared to psychiatric hospitalization) and do not require anticipating acute risk. Pairing predictive modeling with preventative measures is an exciting avenue for efficiently targeting proactive services. The Veterans Health Administration's Recovery Engagement and Coordination for Health–Veterans Enhanced Treatment (REACH VET) program used a predictive model constructed through machine learning to stratify patients by risk-level (VA, 2017). Veterans in the highest risk stratum are then contacted by clinicians who review their treatment plan, further assess clinical risk, and offer support and resources. The REACH VET initiative has demonstrated feasibility (McCarthy et al., 2015) and therefore may represent a beneficial middle ground of prediction and prevention—one which maximizes benefits for those who need care and minimizes the potential harms for those falsely flagged. We have high hopes that the future of suicide prevention will combine such thoughtful uses of predictive modeling with the creation and implementation of effective interventions.

8. Conclusion

Machine learning can be applied to address fundamental questions regarding suicide at multiple levels of investigation, from basic neuroscience research to applied prevention and clinical efforts. It can also be used to produce powerful engines for prediction to serve practical goals within specific medical facilities (e.g., Barak-Corren et al., 2020). Critically, these are distinctly different objectives (Bennett et al., 2019). Machine learning can most effectively supplement and accelerate scientific progress when there is deliberate synthesis between theory- and data-driven exploration. Thus, its application must be tailored to every use case, and it is essential that those building predictive models—and those interpreting the results of predictive models—have a working understanding of machine learning and the various ways in which it can be applied. The goal we all share is the early identification of individuals truly at risk for suicide so that intervention is possible. The most rapid path toward that goal may be one which there is a productive feedback loop between hypothesis, machine learning, and theory development.

Author contributions

The work was conceptualized by Coxand Tucker. The literature was investigated primarily by Moscardini and Tucker. The manuscript was originally drafted by Cox, with revisions and edits by all authors. The submitted manuscript was approved by all authors.

Declaration of Competing Interest

None.

Acknowledgements

This work was supported by startup funds provided by Louisiana State University.

References

Amini, P., Ahmadinia, H., Poorolajal, J., & Moqaddasi Amiri, M. (2016). Evaluating the high risk groups for suicide: A comparison of logistic regression, support vector machine, decision tree and artificial neural network. *Iranian Journal of Public Health, 45*, 1179–1187.

Barak-Corren, Y., Castro, V. M., Nock, M. K., Mandl, K. D., Madsen, E. M., Seiger, A., … Smoller, J. W. (2020). Validation of an electronic health record–based suicide risk prediction modeling approach across multiple health care systems. *JAMA Network Open, 3* (e201262-e62).

Beam, A. L., & Kohane, I. S. (2018). Big data and machine learning in health care. *JAMA, 319*, 1317–1318.

Belsher, B. E., Smolenski, D. J., Pruitt, L. D., Bush, N. E., Beech, E. H., Workman, D. E., … Skopp, N. A. (2019). Prediction models for suicide attempts and deaths: A systematic review and simulation. *JAMA Psychiatry, 76*, 642–651.

Bennett, D., Silverstein, S. M., & Niv, Y. (2019). The two cultures of computational psychiatry. *JAMA Psychiatry, 76*, 563–564.

Bernecker, S. L., Rosellini, A. J., Nock, M. K., Chiu, W. T., Gutierrez, P. M., Hwang, I., … Kessler, R. C. (2018). Improving risk prediction accuracy for new soldiers in the U.S. Army by adding self-report survey data to administrative data. *BMC Psychiatry, 18*, 87.

Burnaev, E., Erofeev, P., & Papanov, A. (2015). Influence of resampling on accuracy of imbalanced classification. In *, 9875. Eighth international conference on machine vision (ICMV 2015)* (p. 987521). International Society for Optics and Photonics.

Burnap, P., Colombo, G., Amery, R., Hodorog, A., & Scourfield, J. (2017). Multi-class machine classification of suicide-related communication on twitter. *Online Social Networks and Media, 2*, 32–44.

Busch, K. A., Fawcett, J., & Jacobs, D. G. (2003). Clinical correlates of inpatient suicide. *The Journal of Clinical Psychiatry., 64*(1), 14–19.

Bzdok, D., & Meyer-Lindenberg, A. (2018). Machine learning for precision psychiatry: Opportunities and challenges. *Biological Psychiatry: Cognitive Neuroscience and Neuroimaging, 3*, 223–230.

Cabitza, F., Rasoini, R., & Gensini, G. F. (2017). Unintended consequences of machine learning in medicine. *JAMA, 318*, 517.

Center for Disease Control. (2018). *National violent death reporting system. 2020*.

Chekroud, A. M., Zotti, R. J., Shehzad, Z., Gueorguieva, R., Johnson, M. K., Trivedi, M. H., … Corlett, P. R. (2016). Cross-trial prediction of treatment outcome in depression: A machine learning approach. *The Lancet Psychiatry, 3*, 243–250.

Chen, R.-B., Chu, C.-H., Yuan, S., & Wu, Y. N. (2016). Bayesian sparse group selection. *Journal of Computational and Graphical Statistics, 25*, 665–683.

Cheng, Q., Li, T. M., Kwok, C.-L., Zhu, T., & Yip, P. S. (2017). Assessing suicide risk and emotional distress in Chinese social media: A text mining and machine learning study. *Journal of Medical Internet Research, 19*, Article e243.

Chung, D. T., Ryan, C. J., Hadzi-Pavlovic, D., Singh, S. P., Stanton, C., & Large, M. M. (2017). Suicide rates after discharge from psychiatric facilities: A systematic review and meta-analysis. *JAMA Psychiatry, 74*, 694–702.

Cohen, A. S., Cox, C. R., Le, T. P., Cowan, T. M., Masucci, M., Strauss, G. P., & Kirkpatric, B. (in press). Using machine learning of computerized vocal expression to measure blunted vocal affect and alogia. *NPJ Schizophrenia.*

Cohen, A. S., Schwartz, E., Le, T., Cowan, T., Cox, C. R., Tucker, R., … Elvevåg, B. (2019). Validating digital phenotyping technologies for clinical use: the critical importance of "resolution". *World Psychiatry, 19*(1), 114–115.

Cox, D. D., & Savoy, R. L. (2003). Functional magnetic resonance imaging (fMRI) "brain reading": Detecting and classifying distributed patterns of fMRI activity in human visual cortex. *NeuroImage, 19*, 261–270.

Davatzikos, C., Ruparel, K., Fan, Y., Shen, D. G., Acharyya, M., Loughead, J. W., … Langleben, D. D. (2005). Classifying spatial patterns of brain activity with machine learning methods: Application to lie detection. *Neuroimage, 28*, 663–668.

Delgadillo, J., & Gonzalez Salas Duhne, P. (2020). Targeted prescription of cognitive–behavioral therapy versus person-centered counseling for depression using a machine learning approach. *Journal of Consulting and Clinical Psychology, 88*, 14.

Figueiredo, M. A. T., & Nowak, R. D. (2016). Ordered weighted ℓ1 regularized regression with strongly correlated covariates: Theoretical aspects. In *Proceedings of the 19th international conference on artificial intelligence and statistics* (pp. 930–938). Cadiz, Spain: PMLR.

Franklin, J. C., Ribeiro, J. D., Fox, K. R., Bentley, K. H., Kleiman, E. M., Huang, X., … Nock, M. K. (2017). Risk factors for suicidal thoughts and behaviors: A meta-analysis of 50 years of research. *Psychological Bulletin, 143*, 187.

Glaros, A. G., & Kline, R. B. (1988). Understanding the accuracy of tests with cutting scores: The sensitivity, specificity, and predictive value model. *Journal of Clinical Psychology, 44*, 1013–1023.

Hanson, S. J., Matsuka, T., & Haxby, J. V. (2004). Combinatorial codes in ventral temporal lobe for object recognition: Haxby (2001) revisited: Is there a "face" area? *Neuroimage, 23*, 156–166.

Hedegaard, H., Curtin, S. C., & Warner, M. (2018). Suicide rates in the United States continue to increase. In , *309*. *NCHS data brief*. Hyattsville, MD: National Center for Health Statistics.

Hill, R. M., Oosterhoff, B., & Kaplow, J. B. (2017). Prospective identification of adolescent suicide ideation using classification tree analysis: Models for community-based screening. *Journal of Consulting and Clinical Psychology, 85*, 702.

Hoerl, A. E., & Kennard, R. W. (1970). Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics, 12*, 55–67.

Hornik, K. (1991). Approximation capabilities of multilayer feedforward networks. *Neural Networks, 4*, 251–257.

Hornik, K., Stinchcombe, M., & White, H. (1989). Multilayer feedforward networks are universal approximators. *Neural Networks, 2*, 359–366.

Jaggi, M. (2014). An equivalence between the lasso and support vector machines. In J. A. K. Suykens, M. Signoretto, & A. Argyriou (Eds.), *Regularization, optimization, kernels, and support vector machines*. Chapman and Hall.

Joiner, T. E., Jr., Conwell, Y., Fitzpatrick, K. K., Witte, T. K., Schmidt, N. B., Berlim, M. T., … Rudd, M. D. (2005). Four studies on how past and current suicidality relate even when" everything but the kitchen sink" is covaried. *Journal of Abnormal Psychology, 114*, 291.

Just, M. A., Pan, L., Cherkassky, V. L., McMakin, D. L., Cha, C., Nock, M. K., & Brent, D. (2017). Machine learning of neural representations of suicide and emotion concepts identifies suicidal youth. *Nature Human Behaviour, 1*, 911–919.

Kamitani, Y., & Tong, F. (2005). Decoding the visual and subjective contents of the human brain. *Nature Neuroscience, 8*, 679–685.

Kapur, N., Hunt, I., Windfuhr, K., Rodway, C., Webb, R., Rahman, M., … Appleby, L. (2013). Psychiatric in-patient care and suicide in England, 1997 to 2008: A longitudinal study. *Psychological Medicine, 43*, 61–71.

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*, 544–551.

Kessler, R. C., van Loo, H. M., Wardenaar, K. J., Bossarte, R. M., Brenner, L. A., Cai, T., … de Jonge, P. (2016). Testing a machine-learning algorithm to predict the persistence and severity of major depressive disorder from baseline self-reports. *Molecular Psychiatry, 21*, 1366–1371.

King, C. A., Arango, A., Kramer, A., Busby, D., Czyz, E., Foster, C. E., & Gillespie, B. W. (2019). Association of the youth-nominated support team intervention for suicidal adolescents with 11-to 14-year mortality outcomes: Secondary analysis of a randomized clinical trial. *JAMA Psychiatry, 76*, 492–498.

Kleiman, E. M., Turner, B. J., Fedor, S., Beale, E. E., Huffman, J. C., & Nock, M. K. (2017). Examination of real-time fluctuations in suicidal ideation and its risk factors: Results from two ecological momentary assessment studies. *Journal of Abnomal Psychology, 126*, 726–738.

Klonsky, E. D., & May, A. M. (2015). The three-step theory (3ST): A new theory of suicide rooted in the "ideation-to-action" framework. *International Journal of Cognitive Therapy, 8*, 114–129.

Klonsky, E. D, May, A. M., & Boaz, Y. (2016). Saffer, Suicide, Suicide Attempts, and Suicidal Ideation. *Annual Review of Clinical Psychology, 12*(1), 307–330.

Klonsky, E. D., Saffer, B. Y., & Bryan, C. J. (2018). Ideation-to-action theories of suicide: A conceptual and empirical update. *Current Opinion in Psychology, 22*, 38–43.

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*, 9776–9789.

Liong, C. Y., & Foo, S. F. (2013). Comparison of linear discriminant analysis and logistic regression for data classification. In , *1522*. *AIP conference proceedings* (pp. 1159–1165).

Madsen, T., Erlangsen, A., & Nordentoft, M. (2017). Risk estimates and risk factors related to psychiatric inpatient suicide—An overview. *International Journal of Environmental Research and Public Health, 14*, 253.

Marquardt, D. W., & Snee, R. D. (1975). Ridge regression in practice. *The American Statistician, 29*, 3–20.

Matsubayashi, T., Ueda, M., & Sawada, Y. (2014). The effect of public awareness campaigns on suicides: Evidence from Nagoya, Japan. *Journal of Affective Disorders, 152*, 526–529.

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*, 1935–1942.

McHugh, C. M., & Large, M. M. (2020). Can machine-learning methods really help predict suicide? *Current Opinion in Psychiatry, 33*(4), 369–374.

Moradi, E., Pepe, A., Gaser, C., Huttunen, H., Tohka, J., & Initiative, A. s. D. N.. (2015). Machine learning framework for early MRI-based Alzheimer's conversion prediction in MCI subjects. *Neuroimage, 104*, 398–412.

Nathoo, F. S., Greenlaw, K., & Lesperance, M. (2016). *Regularization parameter selection for a Bayesian group sparse multi-task regression model with application to imaging genomics. In International Workshop on Pattern Recognition in Neuroimaging*. Trento: Institute of Electrical and Electronics Engineers Inc.

Nock, M. K., Millner, A. J., Joiner, T. E., Gutierrez, P. M., Han, G., Hwang, I., … Kessler, R. C. (2018). Risk factors for the transition from suicide ideation to suicide attempt: Results from the Army study to assess risk and resilience in Servicemembers (Army STARRS). *Journal of Abnormal Psychology, 127*, 139–149.

O'Connor, R. C., Cleare, S., Eschle, S., Wetherall, K., & Kirtley, O. J. (2016). The integrated motivational-volitional model of suicidal behavior: An update. In *The international handbook of suicide prevention* (pp. 220–240).

Paksarian, D., Mojtabai, R., Kotov, R., Cullen, B., Nugent, K. L., & Bromet, E. J. (2014). Perceived trauma during hospitalization and treatment participation among individuals with psychotic disorders. *Psychiatric Services, 65*, 266–269.

Pereira, F., Mitchell, T., & Botvinick, M. (2009). Machine learning classifiers and fMRI: A tutorial overview. *Neuroimage, 45*, S199–S209.

Rao, N. S., Cox, C. R., Nowak, R. D., & Rogers, T. T. (2013). *Sparse overlapping sets lasso for multitask learning and its application to fMRI analysis*.

Rao, N. S., Nowak, R. D., Cox, C. R., & Rogers, T. T. (2016). Classification with the sparse group lasso. *IEEE Transactions on Signal Processing, 64*, 448–463.

Ribeiro, J. D., Franklin, J. C., Fox, K. R., Bentley, K. H., Kleiman, E. M., Chang, B. P., & Nock, M. K. (2016). Letter to the editor: Suicide as a complex classification problem: Machine learning and related techniques can advance suicide prediction - a reply to Roaldset (2016). *Psychological Medicine, 46*, 2009–2010.

Simon, G. E., Johnson, E., Lawrence, J. M., Rossom, R. C., Ahmedani, B., Lynch, F. L., … Shortreed, S. M. (2018). Predicting suicide attempts and suicide deaths following outpatient visits using electronic health records. *The American Journal of Psychiatry, 175*, 951–960.

Simon, N., Friedman, J., Hastie, T., & Tibshirani, R. (2013). A sparse-group lasso. *Journal of Computational and Graphical Statistics, 22*, 231–245.

Stanley, B., & Brown, G. K. (2012). Safety planning intervention: A brief intervention to mitigate suicide risk. *Cognitive and Behavioral Practice, 19*, 256–264.

Stanley, B., Brown, G. K., Brenner, L. A., Galfalvy, H. C., Currier, G. W., Knox, K. L., … Green, K. L. (2018). Comparison of the safety planning intervention with follow-up vs usual care of suicidal patients treated in the emergency department. *JAMA Psychiatry, 75*, 894–900.

Thabtah, F. (2019). Machine learning in autistic spectrum disorder behavioral research: A review and ways forward. *Informatics for Health & Social Care, 44*, 278–297.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B: Methodological, 58*, 267–288.

Torgo, L., Branco, P., Ribeiro, R. P., & Pfahringer, B. (2015). Resampling strategies for regression. *Expert Systems, 32*, 465–476.

Torous, J., Larsen, M. E., Depp, C., Cosco, T. D., Barnett, I., Nock, M. K., & Firth, J. (2018). Smartphones, sensors, and machine learning to advance real-time prediction and interventions for suicide prevention: A review of current progress and next steps. *Current Psychiatry Reports, 20*, 51.

Tran, T., Luo, W., Phung, D., Harvey, R., Berk, M., Kennedy, R. L., & Venkatesh, S. (2014). Risk stratification using data from electronic medical records better predicts suicide risks than clinician assessments. *BMC Psychiatry, 14*, 76.

VA, D. O. V. A. (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: (The Reach Vet press release).

Van Orden, K. A., Witte, T. K., Cukrowicz, K. C., Braithwaite, S. R., Selby, E. A., & Joiner, T. E., Jr. (2010). The interpersonal theory of suicide. *Psychological Review, 117*, 575.

Walsh, C. G., Ribeiro, J. D., & Franklin, J. C. (2017). Predicting risk of suicide attempts over time through machine learning. *Clinical Psychological Science, 5*, 457–469.

Ward-Ciesielski, E. F., & Rizvi, S. L. (2020). *The potential iatrogenic effects of psychiatric hospitalization for suicidal behavior: A critical review and recommendations for research* (p. e12332). Clinical Psychology: Science and Practice.

Weiss, G. M. (2004). Mining with rarity: A unifying framework. *ACM Sigkdd Explorations Newsletter, 6*, 7–19.

World Health Organization. (2017). *Depression and other common mental disorders: Global health estimates*. In: World Health Organization.

Yuan, M., & Lin, Y. (2006). Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society, Series B: Statistical Methodology, 68*, 49–67.

Christopher Cox completed his PhD in cognitive psychology at the University of Wisconsin-Madison, where he helped develop machine learning techniques for the analysis of functional neuroimages. He then joined the Neuroscience and Aphasia Research Unit at the University of Manchester as a post-doctoral researcher to implement these techniques and explore the representation of semantic memory in brain-wide networks. Dr. Cox is currently an assistant professor of Psychology in the Cognition and Brain Science research group at Louisiana State University applying computational approaches to the study of learning and knowledge representation.

Emma Moscardini is a graduate student of clinical psychology in the Mitigation of Suicidal Behaviors Lab under the supervision of Dr. Raymond Tucker. Her work focuses on theoretical models of suicide with the aim of better understanding risk and resiliency

factors. She is also broadly interested in how adversity impacts an individual's experience of meaning in life.

Alex Cohen is a licensed clinical psychologist who focuses on understanding and improving the lives of individuals with serious mental illnesses. His work focuses on adapting biobehavioral technologies, notably automated computerized analysis of natural behavior, for assessing a wide range of clinical issues, including suicidality, depression, psychosis, mania and anxiety. Dr. Cohen is currently a professor of Clinical Psychology at Louisiana State University and adjunct at Pennington Biomedical Research Center and LSU Health Sciences. His clinical team provides clinical services at a variety of outpatient and inpatient clinics in the greater Baton Rouge area.

Raymond Tucker conducts research in the field of suicide prevention with a specific focus on novel contributors to suicide risk and historical and current cultural factors that influence suicide risk and resilience in underrepresented populations (e.g., transgender veterans and ethnic/racial minority adults). Dr. Tucker's research program is also active in online awareness campaigns.