Individual level prediction of emerging suicide events in the pharmacologic treatment of bipolar disorder

Sean X. Luo, MD, PhD^{*1}

Adam Ciarleglio, PhD²

Hanga Galfalvy, PhD¹

Michael Grunebaum, MD¹

Leo Sher, MD³

J. John Mann, MD¹

Maria A. Oquendo, MD⁴

*corresponding author

Xsl2101@columbia.edu

1. New York State Psychiatric Institute and Department of Psychiatry, Columbia University, 1051 Riverside Dr, New York, NY 10032

2. Department of Biostatistics and Bioinformatics, George Washington University Milken Institute School of Public Health, 950 New Hampshire Ave NW, Washington, DC 20052

3. Department of Psychiatry, Icahn School of Medicine at Mount Sinai, 1 Gustave L. Levy Pl, New York, NY 10029

4. Department of Psychiatry, University of Pennsylvania, 3535 Market Street, Philadelphia, PA 19104

Category: Regular research article

Word Count: Abstract : 201 | Body: 3781

Number of References: 52

Number of Figures 1 and Tables 3

Significance Statement

This study aims to find out whether suicide events that occur during the pharmacological treatment of bipolar disorder, a severe psychiatric disorder that is highly associated with suicide behavior, can be predicted. Using existing methods, we developed and compared several predictive models. We showed that these models performed similarly to predictive models of other outcomes, such as treatment efficacy, in unipolar and bipolar depression. This suggests that suicide events during bipolar disorder may be a feasible target for individualized interventions in the future.

Abstract

Background

Patients with bipolar disorder have a high lifetime risk of suicide. Predicting, preventing and managing suicidal behavior are major goals in clinical practice. Changes in suicidal thoughts and behavior are common in the course of treatment of bipolar disorder.

Methods

Using a dataset from a randomized clinical trial of bipolar disorder treatment (N=98), we tested predictors of future suicidal behavior identified through a review of literature and applied marginal variable selection and machine learning methods. The performance of the models was assessed using the optimism-adjusted C statistic.

Results

Number of prior hospitalizations, number of prior suicide attempts, current employment status and Hamilton Depression Scale were identified as predictors and a simple logistic regression model was constructed. This model was compared with a model incorporating interactions with treatment group assignment, and more complex variable selection methods (LASSO and Survival Trees). The best performing models had average optimism-adjusted C-statistics of 0.67 (main effects only) and 0.69 (Survival Trees). Incorporating medication group did not improve prediction performance of the models.

Conclusions

These results suggest that models with a few predictors may yield a clinically meaningful way to stratify risk of emerging suicide events in patients who are undergoing pharmacologic treatment for bipolar disorder.

Keywords

Bipolar disorder; lithium; valproate; predictive modeling; machine learning; suicide

Introduction

Bipolar disorder is associated with elevated risks for suicide attempt and suicide death (Chen and Dilsaver), with a commonly cited lifetime risk for both at 15-25% (Oquendo et al., 2010).

Several lines of indirect evidence suggest that on a population level, lithium may be the preferred agent for preventing suicidal behavior compared to other mood stabilizers (Cipriani et al., 2005; Baldessarini et al., 2006), although not all studies agree (Lauterbach et al., 2008), and discontinuation of lithium, especially rapid discontinuation, may increase risk for suicidal behavior (Muller-Oerlinghausen et al., 1992; Baldessarini et al., 1996; Baldessarini et al., 1999). Relatively little is known about the effect on suicidal behavior of anticonvulsant mood stabilizers such as valproate (Goodwin) and lamotrigine relative to lithium (Bowden et al., 2003; Calabrese et al., 2003). A double blind, parallel-group active comparison study (Oquendo et al., 2011) was conducted to compare lithium and valproate and did not show a relative advantage for lithium in preventing suicide attempt or suicide event, defined as suicidal ideation with a plan and requiring clinical intervention. While the study may have lacked statistical power due to the modest sample size, an alternative and additional explanation is that there may exist heterogeneities in treatment outcome: lithium might work better for some patients, while valproate might work better for others. Testing for moderator-treatment interaction, however, is also limited by the small sample size.

We sought to approach this problem by modeling treatment response heterogeneity directly. A clearer understanding of treatment response heterogeneity may allow researchers to develop individualized treatment rules to tailor treatment on an individual level (Qian and Murphy, 2011),

and such strategies have shown promise in other areas of medicine such as oncology (Friedman et al., 2015) and cardiology (Antman and Loscalzo, 2016). Similar work has been emerging recently in psychiatry for major depression (Chekroud et al., 2016) autism (Abraham et al., 2017), and suicidal risk stratification using non-pharmacological datasets (Goldstein et al., 1991). A variety of treatment by baseline covariate interactions have been reported in depression: serotonergic antidepressants were reported to be more effective in patients with more severe depression (Fournier et al.), and, in particular, were more effective compared with a noradrenergic antidepressant in treating suicidal thoughts and behavior as baseline suicidal ideation severity increased (Grunebaum et al., 2012). While prospective observational studies identified a number of significant single predictors in major mood disorders (Oquendo et al., 2004; Sokero et al., 2005; Galfalvy et al., 2006) for suicide attempts and death, multi-predictor individual-level models based on large, and presumably heterogeneous datasets have been less successful (Goldstein et al., 1991; Marangell et al., 2006). Recently, new models using machine learning techniques and larger observational datasets, such as the Army Study To Assess Risk & Resilience in Servicemembers (STARSS) dataset, allow for more accurate risk stratification for individual level prediction of suicide risk after hospitalization (Kessler et al., 2015) and through outpatient mental health visits (Kessler et al., 2017b). These predictive models have also been built for electronic medical records (Kessler et al., 2017a). Nevertheless, these approaches have not been applied to treatment datasets, and therefore we do not know whether the predictive information can allow for tailoring of treatment, especially because so few clinical trials specifically targeted changes in suicidal thoughts and behavior as the primary outcome.

In the present study, we applied existing methodological approaches by building multi-predictor, predictive survival models of suicide event using a treatment dataset containing data from individuals undergoing pharmacological treatment of bipolar disorder. We sought to build models with both treatment group assignment and patient characteristics (i.e. models incorporating covariate by treatment interactions) have better predictive performance compared to models with only patient characteristics. This addresses not just the impact of treatment group on outcome, but also whether there could be improvement of outcome when treatment choice is tailored to individuals' characteristics. The goal of the analysis is to provide a foundation that future studies targeting emerging suicidal ideations and behavior can also be plausibly designed to be more risk adaptive (Tomlinson et al., 2020).

Material and Methods

Trial Methodology Summary

This study was based on a pragmatic, parallel-group 2.5-year double-blinded trial for patients with bipolar disorder and past suicide attempts (N=98) (Oquendo et al., 2011). All participants provided written informed consent as approved by the New York State Psychiatric Institute Institutional Review Board. The trial enrolled a broadly inclusive group of patients with bipolar disorder (bipolar I or II disorder or bipolar disorder not otherwise specified, DSM-IV criteria); the participants were in a depressive or mixed episode; had at least one past suicide attempt; and between 18 to 75 years of age. Patients were also required to have had no history of nonresponse to adequate dosages of either lithium or valproate in the past 2 years. They were randomized at the start of the trial to either lithium or valproate with dosages optimized by an un-blinded

physician based on blood levels. The design closely simulated clinical practice such that adjunct and second line treatment options followed existing guidelines for treatment of affective episodes in bipolar disorder. For further details on trial methodology, see (Oquendo et al., 2011).

Baseline characteristics

Baseline psychopathology, demographic and patient characteristics were assessed using the Structured Clinical Interview for DSM-IV Axis I and II Disorders (SCID) as well as other structured inventories. Although patients were followed through a 2.5-year period, in the present study we only examined baseline predictors in developing our predictive models, based on the rationale that while early treatment efficacy might correlate with a good outcome, a predictive model based on only baseline characteristics would be more clinically useful.

Outcome Measures

In the original trial, the primary outcome measures were time to suicide completion, time to suicide attempt and time to suicide event. Because there were no completed suicides in either group, and nonfatal suicide attempts were also rare, we examined time to suicide event as the prediction outcome. Suicide event was defined as either a suicide attempt (a potentially self-injurious behavior with some intent to end one's life), or an episode of suicidal ideation with a plan, as assessed using a standardized inventory (Scale for Suicide Ideation), that required a change in treatment, such as addition of a rescue medication or hospitalization. This outcome was one of the primary outcome measures reported in the original study (Oquendo et al., 2011).

Outcome measures were considered to have been at least partially observed for 94 subjects, with some censoring; the other 4 subjects did not return for follow-up visits and were not included in the subsequent analysis. Our choice of suicide event as the primary outcome rests on this outcome being an action response measure, whereas suicide attempts may be suppressed during a pragmatic clinical trial, when clinicians immediately intervene if a suicide attempt may be imminent.

Statistical Analysis

The main goal of our modeling effort is to predict the time to suicide event for an individual patient given the treatment group assignment and other baseline patient characteristics. We used a mixed domain-knowledge (using existing literature in the suicide literature) and data-driven (depending on only the data to provide information on predictors) approach using a broad set of baseline covariates as predictors, similar to the approach proposed in (Kessler et al., 2015; Kessler et al., 2017b). However, unlike those studies, instead of screening all or most of the covariates recorded during the trial for associations with the outcome of interest and then empirically conducting model selection, we surveyed the existing literature to identify predictor covariates that had a theoretical basis for possibly influencing future suicide risk in bipolar disorder.

For each covariate, anywhere between 0 and 17% values were missing. We employed multiple imputation using the MICE package to impute missing covariates (Ian R White et al., 2011). Since we had a time-to-event outcome, we used the Nelson-Aalen estimator for the cumulative hazard of the survival time (rather than log of survival time) in the imputation model (I. R. White

and Royston, 2009) along with all other clinical predictors under consideration. To account for potential interactions between covariates, we used classification and regression tree (CART) imputation (Doove et al., 2014) allowing all predictor variables including treatment group, event status, and the Nelson-Aalen estimator for the cumulative hazard. Using this procedure, we created 5 imputed data sets for analysis.

We constructed the models in the following stages: first we used the APA Guideline for Evaluation of Suicidal Behavior (Fawcett et al., 2009) and listed all the predictors that were proposed in the guideline as a relevant risk predictor during clinical assessment; we then conducted a focused literature search and included proposed predictors that appeared in at least an additional individual research article as a single predictor of suicide risk for any psychiatric condition (citation see Supplementary Material); finally, we cross-matched the proposed predictors to the trial dataset and included those predictors that corresponded to one or more of the items in the baseline assessments. Supplementary material provides definition of the covariates, evidence for inclusion, and the relevant citations (Supplementary Table 1).

In the second stage, we performed marginal screening to estimate the relationship between the outcome and each predictor individually. This allowed for comparing the predictors in single predictor analysis. We used Cox proportional hazard models to estimate the effect of each covariate and the effect of treatment on the final outcome (Table 1A, *single predictor models*) to explore potential predictor variables that have P-values less than 0.05 without specific adjustments for multiple comparisons. For each of the models in this stage and all subsequent

stages, we fit a model to each of the five imputed data sets and pooled the estimates using Rubin's rules (Marshall et al., 2009). We then built models with three additional sets of predictors: treatment group assignment of lithium vs. valproate, the predictor listed, and the predictor by treatment interaction terms (Table 1 B, *interaction models*). To evaluate models from the prediction point of view, we used a method proposed by Smith et al. (Smith et al., 2014) to obtain the optimism-adjusted Harrell's C-statistic as the measure of the predictive performance of the single predictor model. The C-statistic is a commonly used statistical metric of discrimination for individual-level prediction models. In binary prediction models, the C-statistic is the area under the curve (AUC) for a Receiver Operating Characteristic (ROC), and it varies from 0.5 (prediction by chance only) to 1 (perfect accuracy). Harrell's C is a generalization of this concept for survival data, that also varies from 0.5 to 1, and can be thought of as the average AUC for a time-dependent ROC curve, for classifying survival in time.

In the final stage, we estimated the performance of different multi-predictor models for time to suicide event on each of the 5 imputed data sets and again used optimism-adjusted C-statistic as a performance measure:

MODEL 1: Cox model with main effects of predictors only. In this model we selected only those predictors with optimism-adjusted C-values ≥ 0.6 in the *single predictor models*.

MODEL 2: Cox model with main effect and predictor by treatment interactions with optimism-adjusted C-values ≥ 0.6 in either the *single predictor or interaction models*.

MODEL 3: We used LASSO with tuning parameters selected via 10-fold Cross Validation as an data-adaptive method for variable selection of all covariates with only main effect (Tibshirani, 1997). LASSO is a well-known algorithm that exploits a procedure known as L-1 minimization that penalizes the size of the coefficients in a regression model, and therefore typically selects for variables that have relatively strong predictive ability.

MODEL 4: We used LASSO fit as described for Model 3, but for variable selection for both main effect and predictor by treatment interaction terms.

MODEL 5: We employed a Survival Trees strategy (Taylor, 2011). Survival trees are a popular alternative to parametric and semiparametric models. They offer substantial flexibility in the construction of predictive models that allow the data, instead of the modeler, to identify relevant interactions between any of the variables being considered. In constructing survival trees, variable selection is automatic, and the fitted tree can be used to group subjects according to their survival behavior based on relevant covariates (Bou-Hamad et al., 2011). There is no obvious or agreed upon way to pool survival trees from multiply imputed data sets, and a general principle is to look for similarities on multiple trees to obtain a “consistent tree”. To demonstrate this, we plotted a Survival Tree with the corresponding number of events over time in each node (Fig. 1). We defined Variable Importance of a predictor as the mean rank for a predictor in all the Survival Trees obtained in imputed data sets (larger is better), and then ranked the variables from

most important on average to least important. We then calculated the average C-statistic from all imputed datasets.

All analyses were carried out using R Statistical Package (R version 2.12.1 Copyright (C) 2010 The R Foundation for Statistical Computing).

Results

Variable Selection and Marginal Screening

Using the methods outlined, we screened 24 predictors using the *single predictor (main effect only)* and *interaction with treatment group* models (Table 1A/B). We found that four factors reached an arbitrary threshold of optimism-adjusted C of 0.6 (number of hospitalizations, number of suicide attempts, 17-item Hamilton Depression Rating Scale (Ham17), Current Employment (Yes/No)). Number of prior hospitalizations, number of past suicide attempts and current employment status reached the typical p-value threshold of 0.05 based on their corresponding single predictor models. On the other hand, marital status and age appeared to show an interaction effect detected through a conventional p-value threshold of 0.05, but interaction model of neither predictor yielded an optimism adjusted C-statistic of 0.6 or greater, suggesting that the effect sizes of these interactions are insufficient for prediction. While both number of hospitalization and number of suicide attempts reached statistical significance in the *single predictor models*, the effect sizes (Hazard Ratio) were relatively modest (1.27 and 1.37 respectively).

Predictive Model Performance

In model 1, we included the four predictors identified in the marginal screening stage. This multi-predictor Cox model yielded an apparent C of 0.697 and an optimism-adjusted C of 0.671, while 95% bootstrap confidence intervals (CIs) in each of the 5 imputed data sets showed that the smallest lower bound across those CIs was 0.571 and the largest upper bound was 0.800. Adding additional interaction terms (model 2) to this multi-predictor model did not enhance the predictive performance (optimism-adjusted C = 0.653, 95% CI smallest lower bound of 0.579 and largest upper bound of 0.795). In model 3, we used LASSO as a strategy to conduct automatic variable selection of all 24 predictors and obtained a model with an apparent C of 0.719 and an optimism-adjusted C of 0.62. Inclusion of the interaction terms did not appear to improve the performance of the LASSO based models. All of the predictors that appeared in the marginal screening strategy also were selected by the LASSO models.

Survival Tree Modeling

We examined characteristics of the individual trees from each imputed dataset to look for similarities. In order to do so, we plotted survival trees derived from four imputed data sets (Fig. 1) for demonstration purposes. For each tree, we also computed the Variable Importance using recursive partitioning for each variable (Therneau et al., 2015), a classification process that repeatedly splits the sample into sub-samples based on dichotomous independent variables. Based on these importance measures, we assigned a rank to each variable and computed the mean rank for each variable across the 5 imputed data sets. Table 3 shows the top 10 variables in

order of importance from most to least important corresponding to their average rank in the 5 imputed datasets (larger rank more important), providing an indicator of the most important predictors selected by the models. As with the other models, we computed the apparent and optimism-adjusted C value for each survival tree (one for each imputed data set) then took the average. The average apparent C-value is 0.815 and the optimism-adjusted C-value is 0.690.

Discussion

In this study, we built predictive models leveraging data from a randomized controlled trial of lithium versus valproate in treating patients with bipolar disorder who had a high propensity for suicidal behavior (Oquendo et al., 2011). In the original intention-to-treat analysis, no difference in treatment efficacy between lithium and valproate was detected with respect to time to suicide event. Using a predictive modeling approach, our analysis revealed two findings. First, we identified some clinically relevant characteristics that were predictors of outcome. In particular, we found that number of hospitalizations and number of previous suicide attempts, both commonly queried variables in clinical practice, were correlated with time to suicide events. We detected a significant predictor-treatment interaction for marital status and age, though incorporating these variables did not improve predictive performance. Second, we also showed we can predict suicide event with an optimism-adjusted C-value of 0.671 using 4 predictors in a Cox proportional hazards model. The use of more complex modeling, including treatment by baseline interactions, did not demonstrate significant improvement in predictive performance.

These results have several implications, both methodologically and clinically. Previously, predictive modeling work for suicidal behavior relied primarily on observational datasets. Our results suggest that similar modeling strategies can be used for clinical trial datasets for bipolar disorder. The performance of our model is similar to work (Chekroud et al., 2016) in unipolar depression treatment response prediction, with similar individual level predictive performance: cross-validated C-statistic between 0.5 to 0.65 for various medications. A recent meta-analysis suggests that prediction models for bipolar and unipolar depression treatment response using clinical phenomenological predictors typically achieve a c-statistic between 0.65 and 0.75 (Lee et al., 2018). This suggests that while suicide events are in general a rare outcome, in enriched samples such as the one presented here, it may be feasible to use individual level predictive modeling as a technique to infer risk at the baseline, and design adaptive treatment trials specifically targeting suicidal behavior as an outcome.

Our results demonstrated that marginal variable selection, i.e. a data-driven approach based on single predictor models, can perform as well as or better than more complex variable selection methods, such as using LASSO, for sample sizes typical in suicide intervention studies. This is a consistent finding in application of machine learning to clinical trial datasets in other areas of psychiatry such as PTSD (Galatzer-Levy et al., 2014) and addiction (Luo et al., 2015). These examples of psychiatry treatment research, especially the single center studies, appear to lack sufficient sample size to constrain predictive models with large numbers of predictors. Failure to detect substantial improvement using complex non-parametric or non-linear modeling is also a common finding. For example, random forest models using 47 predictors and a simplified model using 15 predictors performed similarly (75% vs. 70%) in a study of prediction of treatment

resistant depression (Kautzky et al., 2018). On the other hand, some datasets, such as those obtained from neuroimaging and genetics, may especially benefit from more complex feature modeling (Lee et al., 2018), and combining data sources may improve predictive performance (see recent reports from the EMBARC study (Trivedi et al., 2016; Petkova et al., 2017; Webb et al., 2018)).

Our results also raise several issues specific to predictive modeling for suicidal behavior. A recent meta-analysis reports that despite many years of research, individual risk factors do not provide enough information on their own to predict suicide, and prediction of suicide events over a long time period is disappointingly ineffective (Franklin et al., 2017). Several proposed solutions, including a more comprehensive survey of individual level characteristics through analyses of medical records, have yielded some early promising findings: a multi-stage approach for example, can detect future suicide attempts in U.S. Army soldiers who had no previous known suicidal ideation (Bernecker et al., 2018). Studies that involve electronic medical records often reach higher C-statistics (Kessler et al., 2017a), and these models are particularly useful in a large system in leveraging numerous alerts be sent to clinicians at lower cost for a heterogenous population. Our results suggest an additional approach: we may wish to identify those who are already diagnosed or being treated with a psychiatric disorder that make a suicide event much more likely. These enriched samples might allow us to treat these patients differently based on risk stratification. This is a secondary prevention strategy that is common in other fields of medicine, with corresponding risk prediction models and scores: e.g. Thrombolysis In Myocardial Infarction (TIMI) risk score for prediction of ischemic events for individuals who already have unstable angina; and the CHADS2 score (congestive heart failure, hypertension,

age ≥ 75 years, diabetes mellitus, stroke [double weight]) for prediction of stroke for individuals who already have atrial fibrillation. For suicide outcomes, we believe the next step is to similarly construct more definitive predictive scores using larger, multi-center effectiveness trial datasets for similar high-risk psychiatric conditions (i.e. schizophrenia, bipolar disorder and borderline personality disorder). Secondly, an interesting unresolved question is whether results (and simple risk scores) from randomized trials generalize to observational studies of a much larger size, especially as many individuals followed in observational studies are treated with psychopharmacological agents—often they are not (Blanco et al., 2008). Thirdly, the potential to predict outcome and personalize treatment for patients with suicidal bipolar disorder may improve future trial design: we may therefore consider evaluating a specific suicide-targeted treatment program for individuals who are at high risk of suicide event as predicted by these models.

There are a number of limitations of this study. First, the study included only those bipolar patients with a prior suicide attempt, thus it may not generalize to the prediction of a first attempt. We deliberately limited our predictors to baseline characteristics, as they were the most applicable in clinical practice as a means to stratify patients at initial diagnosis. However, predictors that partially indicated treatment success, such as adherence records and early symptomatic resolutions, might also be useful predictors of eventual treatment outcome. Fitting and validating predictive models with a small dataset has a number of technical limitations (Braga-Neto and Dougherty, 2004; Isaksson et al., 2008), and may be prone to false positives and generate an overly confident estimate of predictive performance (Button et al., 2013)—this is partially addressed by using the optimism-adjusted C metric (Smith et al., 2014), though this

method still underestimates true error rate, and more importantly, variance of the error rate. The promising conclusions of this study therefore would require external validation in a larger sample study. Finally, adoption of externally validated predictive models in clinical practice would involve centralized resources, such as the Memorial-Sloan Kettering Prediction Tools (Vickers, 2011), that have been widely used in other medical specialties but are not yet available in psychiatry.

Acknowledgements

None

Statement of Conflict of Interest

Dr. Oquendo receives royalties from the Research Foundation for Mental Hygiene for the commercial use of the Columbia Suicide Severity Rating Scale and owns shares in Mantra, Inc. She serves as an advisor to Alkermes and Fundacion Jimenez Diaz (Madrid). Her family owns stock in Bristol Myers Squibb.

Dr. Mann receives royalties from the Research Foundation for Mental Hygiene for commercial use of the C-SSRS.

The other authors have no disclosures

REFERENCES

- Abraham A, Milham MP, Di Martino A, Craddock RC, Samaras D, Thirion B, Varoquaux G (2017) Deriving reproducible biomarkers from multi-site resting-state data: An Autism-based example. *Neuroimage* 147:736-745.
- Antman EM, Loscalzo J (2016) Precision medicine in cardiology. *Nat Rev Cardiol* 13:591-602.
- Baldessarini RJ, Tondo L, Hennen J (1999) Effects of lithium treatment and its discontinuation on suicidal behavior in bipolar manic-depressive disorders. *J Clin Psychiatry* 60 Suppl 2:77-84; discussion 111-116.
- Baldessarini RJ, Tondo L, Faedda GL, Suppes TR, Floris G, Rudas N (1996) Effects of the rate of discontinuing lithium maintenance treatment in bipolar disorders. *J Clin Psychiatry* 57:441-448.
- Baldessarini RJ, Tondo L, Davis P, Pompili M, Goodwin FK, Hennen J (2006) Decreased risk of suicides and attempts during long-term lithium treatment: a meta-analytic review. *Bipolar Disord* 8:625-639.
- Bernecker SL, Zuromski KL, Gutierrez PM, Joiner TE, King AJ, Liu H, Nock MK, Sampson NA, Zaslavsky AM, Stein MB, Ursano RJ, Kessler RC (2018) Predicting suicide attempts among soldiers who deny suicidal ideation in the Army Study to Assess Risk and Resilience in Servicemembers (Army STARRS). *Behav Res Ther*.
- Blanco C, Olfson M, Okuda M, Nunes EV, Liu SM, Hasin DS (2008) Generalizability of clinical trials for alcohol dependence to community samples. *Drug Alcohol Depend* 98:123-128.
- Bou-Hamad I, Larocque D, Ben-Ameur H (2011) A review of survival trees. *Statistics Surveys* 5:44-71.
- Bowden CL, Calabrese JR, Sachs G, Yatham LN, Asghar SA, Hompland M, Montgomery P, Earl N, Smoot TM, DeVeaugh-Geiss J, Lamictal 606 Study G (2003) A placebo-controlled 18-month trial of lamotrigine and lithium maintenance treatment in recently manic or hypomanic patients with bipolar I disorder. *Arch Gen Psychiatry* 60:392-400.
- Braga-Neto UM, Dougherty ER (2004) Is cross-validation valid for small-sample microarray classification? *Bioinformatics* 20:374-380.
- Button KS, Ioannidis JPA, Mokrysz C, Nosek BA, Flint J, Robinson ESJ, Munafò MR (2013) Power failure: why small sample size undermines the reliability of neuroscience. *Nature Reviews Neuroscience* 14:365-376.
- Calabrese JR, Bowden CL, Sachs G, Yatham LN, Behnke K, Mehtonen OP, Montgomery P, Ascher J, Paska W, Earl N, DeVeaugh-Geiss J, Lamictal 605 Study G (2003) A placebo-controlled 18-month trial of lamotrigine and lithium maintenance treatment in recently depressed patients with bipolar I disorder. *J Clin Psychiatry* 64:1013-1024.

Chekroud AM, Zotti RJ, Shehzad Z, Gueorguieva R, Johnson MK, Trivedi MH, Cannon TD, Krystal JH, Corlett PR (2016) Cross-trial prediction of treatment outcome in depression: a machine learning approach. *Lancet Psychiatry* 3:243-250.

Chen YW, Dilsaver SC (1996) Lifetime rates of suicide attempts among subjects with bipolar and unipolar disorders relative to subjects with other Axis I disorders. *Biol Psychiatry* 39:896-899.

Cipriani A, Pretty H, Hawton K, Geddes JR (2005) Lithium in the prevention of suicidal behavior and all-cause mortality in patients with mood disorders: a systematic review of randomized trials. *Am J Psychiatry* 162:1805-1819.

Doove LL, Van Buuren S, Dusseldorp E (2014) Recursive partitioning for missing data imputation in the presence of interaction effects. *Computational Statistics & Data Analysis* 72:92-104.

Fawcett JA, Baldessarini RJ, Coryell WH, Silverman MM, Stein DJ (2009) Definition and management of suicidality in psychiatric patients. *J Clin Psychiatry* 70:e38.

Fournier JC, DeRubeis RJ, Hollon SD, Dimidjian S, Amsterdam JD, Shelton RC, Fawcett J (2010) Antidepressant drug effects and depression severity: a patient-level meta-analysis. *JAMA* 303:47-53.

Franklin JC, Ribeiro JD, Fox KR, Bentley KH, Kleiman EM, Huang X, Musacchio KM, Jaroszewski AC, Chang BP, Nock MK (2017) Risk factors for suicidal thoughts and behaviors: A meta-analysis of 50 years of research. *Psychol Bull* 143:187-232.

Friedman AA, Letai A, Fisher DE, Flaherty KT (2015) Precision medicine for cancer with next-generation functional diagnostics. *Nat Rev Cancer* 15:747-756.

Galatzer-Levy IR, Karstoft KI, Statnikov A, Shalev AY (2014) Quantitative forecasting of PTSD from early trauma responses: a Machine Learning application. *J Psychiatr Res* 59:68-76.

Galfalvy H, Oquendo MA, Carballo JJ, Sher L, Grunebaum MF, Burke A, Mann JJ (2006) Clinical predictors of suicidal acts after major depression in bipolar disorder: a prospective study. *Bipolar Disord* 8:586-595.

Goldstein RB, Black DW, Nasrallah A, Winokur G (1991) The prediction of suicide. Sensitivity, specificity, and predictive value of a multivariate model applied to suicide among 1906 patients with affective disorders. *Arch Gen Psychiatry* 48:418-422.

Goodwin FK (1999) Anticonvulsant therapy and suicide risk in affective disorders. *J Clin Psychiatry* 60 Suppl 2:89-93; discussion 111-116.

Grunebaum MF, Ellis SP, Duan N, Burke AK, Oquendo MA, John Mann J (2012) Pilot randomized clinical trial of an SSRI vs bupropion: effects on suicidal behavior, ideation, and mood in major depression. *Neuropsychopharmacology* 37:697-706.

Isaksson A, Wallman M, Goransson H, Gustafsson MG (2008) Cross-validation and bootstrapping are unreliable in small sample classification. *Pattern Recognition Letters* 29:1960-1965.

Kautzky A, Dold M, Bartova L, Spies M, Vanicek T, Souery D, Montgomery S, Mendlewicz J, Zohar J, Fabbri C, Serretti A, Lanzenberger R, Kasper S (2018) Refining Prediction in Treatment-Resistant Depression: Results of Machine Learning Analyses in the TRD III Sample. *J Clin Psychiatry* 79.

Kessler RC, Hwang I, Hoffmire CA, McCarthy JF, Petukhova MV, Rosellini AJ, Sampson NA, Schneider AL, Bradley PA, Katz IR, Thompson C, Bossarte RM (2017a) Developing a practical suicide risk prediction model for targeting high-risk patients in the Veterans health Administration. *Int J Methods Psychiatr Res* 26.

Kessler RC, Stein MB, Petukhova MV, Bliese P, Bossarte RM, Bromet EJ, Fullerton CS, Gilman SE, Ivany C, Lewandowski-Romps L, Millikan Bell A, Naifeh JA, Nock MK, Reis BY, Rosellini AJ, Sampson NA, Zaslavsky AM, Ursano RJ, Army SC (2017b) Predicting suicides after outpatient mental health visits in the Army Study to Assess Risk and Resilience in Servicemembers (Army STARRS). *Mol Psychiatry* 22:544-551.

Kessler RC et al. (2015) Predicting suicides after psychiatric hospitalization in US Army soldiers: the Army Study To Assess Risk and rEsilience in Servicemembers (Army STARRS). *JAMA Psychiatry* 72:49-57.

Lauterbach E, Felber W, Muller-Oerlinghausen B, Ahrens B, Bronisch T, Meyer T, Kilb B, Lewitzka U, Hawellek B, Quante A, Richter K, Broocks A, Hohagen F (2008) Adjunctive lithium treatment in the prevention of suicidal behaviour in depressive disorders: a randomised, placebo-controlled, 1-year trial. *Acta Psychiatr Scand* 118:469-479.

Lee Y, Ragguett RM, Mansur RB, Boutilier JJ, Rosenblat JD, Trevizol A, Brietzke E, Lin K, Pan Z, Subramaniapillai M, Chan TCY, Fus D, Park C, Musial N, Zuckerman H, Chen VC, Ho R, Rong C, McIntyre RS (2018) Applications of machine learning algorithms to predict therapeutic outcomes in depression: A meta-analysis and systematic review. *J Affect Disord* 241:519-532.

Luo SX, Covey LS, Hu MC, Levin FR, Nunes EV, Winhusen TM (2015) Toward personalized smoking-cessation treatment: Using a predictive modeling approach to guide decisions regarding stimulant medication treatment of attention-deficit/hyperactivity disorder (ADHD) in smokers. *Am J Addict* 24:348-356.

Marangell LB, Bauer MS, Dennehy EB, Wisniewski SR, Allen MH, Miklowitz DJ, Oquendo MA, Frank E, Perlis RH, Martinez JM, Fagioli A, Otto MW, Chessick CA, Zboyan HA, Miyahara S, Sachs G, Thase ME (2006) Prospective predictors of suicide and suicide attempts in 1,556 patients with bipolar disorders followed for up to 2 years. *Bipolar Disord* 8:566-575.

Marshall A, Altman DG, Holder RL, Royston P (2009) Combining estimates of interest in prognostic modelling studies after multiple imputation: current practice and guidelines. *BMC Med Res Methodol* 9:57.

Muller-Oerlinghausen B, Muser-Causemann B, Volk J (1992) Suicides and parasuicides in a high-risk patient group on and off lithium long-term medication. *J Affect Disord* 25:261-269.

Oquendo MA, Currier D, Liu SM, Hasin DS, Grant BF, Blanco C (2010) Increased risk for suicidal behavior in comorbid bipolar disorder and alcohol use disorders: results from the National Epidemiologic Survey on Alcohol and Related Conditions (NESARC). *J Clin Psychiatry* 71:902-909.

Oquendo MA, Galfalvy H, Russo S, Ellis SP, Grunebaum MF, Burke A, Mann JJ (2004) Prospective study of clinical predictors of suicidal acts after a major depressive episode in patients with major depressive disorder or bipolar disorder. *Am J Psychiatry* 161:1433-1441.

Oquendo MA, Galfalvy HC, Currier D, Grunebaum MF, Sher L, Sullivan GM, Burke AK, Harkavy-Friedman J, Sublette ME, Parsey RV, Mann JJ (2011) Treatment of suicide attempters with bipolar disorder: a randomized clinical trial comparing lithium and valproate in the prevention of suicidal behavior. *Am J Psychiatry* 168:1050-1056.

Petkova E, Ogden RT, Tarpey T, Ciarleglio A, Jiang B, Su Z, Carmody T, Adams P, Kraemer HC, Grannemann BD, Oquendo MA, Parsey R, Weissman M, McGrath PJ, Fava M, Trivedi MH (2017) Statistical Analysis Plan for Stage 1 EMBARC (Establishing Moderators and Biosignatures of Antidepressant Response for Clinical Care) Study. *Contemp Clin Trials Commun* 6:22-30.

Qian M, Murphy SA (2011) Performance Guarantees for Individualized Treatment Rules. *Ann Stat* 39:1180-1210.

Smith GC, Seaman SR, Wood AM, Royston P, White IR (2014) Correcting for optimistic prediction in small data sets. *Am J Epidemiol* 180:318-324.

Sokero TP, Melartin TK, Rytala HJ, Leskela US, Lestela-Mielonen PS, Isometsa ET (2005) Prospective study of risk factors for attempted suicide among patients with DSM-IV major depressive disorder. *Br J Psychiatry* 186:314-318.

Taylor JM (2011) Random Survival Forests. *J Thorac Oncol* 6:1974-1975.

Therneau T, Atkinson B, Ripley BJRpv (2015) rpart: Recursive partitioning and regression trees. 4:1-9.

Tibshirani R (1997) The lasso method for variable selection in the Cox model. *Stat Med* 16:385-395.

Tomlinson A, Furukawa TA, Efthimiou O, Salanti G, De Crescenzo F, Singh I, Cipriani A (2020) Personalise antidepressant treatment for unipolar depression combining individual choices, risks and big data (PETRUSHKA): rationale and protocol. *Evid Based Ment Health* 23:52-56.

Trivedi MH et al. (2016) Establishing moderators and biosignatures of antidepressant response in clinical care (EMBARC): Rationale and design. *J Psychiatr Res* 78:11-23.

Vickers AJ (2011) Prediction models in cancer care. *CA Cancer J Clin*.

Webb CA et al. (2018) Personalized prediction of antidepressant v. placebo response: evidence from the EMBARC study. *Psychol Med*:1-10.

White IR, Royston P (2009) Imputing missing covariate values for the Cox model. *Stat Med* 28:1982-1998.

White IR, Royston P, Wood AMJSim (2011) Multiple imputation using chained equations: issues and guidance for practice. 30:377-399.

Funding

This work was supported by the National Institute of Health (grant number 5K23DA042136).

Figure Legend

Figure 1. **Survival Tree Analyses.** Illustrative Survival Trees delineating various predictor variables as nodes (numhosp – Number of Hospitalizations, rfltot – Reason For Living (Total), fdsuia - Total definite suicide attempts in first degree relatives, bisttot1 – Barratt Impulsivity Scale (attentional), bisttot2 – Barratt Impulsivity Scale (Motor), bisttot3 - Barratt Impulsivity Scale (non-planning), bistoton—Barratt Impulsivity Scale (Total), curempl – Current Employment (1-Yes/0-No), smoke – Tobacco smoking (1-Yes/0-No)) and classification of differential risk groups. Top number in each box is the number of suicide events/time in the node; middle fraction shows number of suicide events out of number of subjects in the node; third number gives the % of participants from the whole sample in that node who had a suicide event. Each tree results in one imputation. Color of the bars indicate the proportion of participants who were partitioned into the branch, with darker color indicating a higher proportion.

Figure 1. Survival Tree Analyses. Illustrative Survival Trees delineating various predictor variables as nodes (numhosp – Number of Hospitalizations, rfltot – Reason For Living (Total), fdsuia - Total definite suicide attempts in first degree relatives, bisttot1 – Barratt Impulsivity Scale (attentional), bisttot2 – Barratt Impulsivity Scale (Motor), bisttot3 - Barratt Impulsivity Scale (non-planning), bisttotn—Barratt Impulsivity Scale (Total), curempl – Current Employment (1-Yes/0-No), smoke – Tobacco smoking (1-Yes/0-No)) and classification of differential risk groups. Top number in each box is the number of suicide events/time in the node; middle fraction shows number of suicide events out of number of subjects in the node; third number gives the % of participants from the whole sample in that node who had a suicide event. Each tree results in one imputation. Color of the bars indicate the proportion of participants who were partitioned into the branch, with darker color indicating a higher proportion.

Table 1. Marginal Screening for Predictively Informative Covariates. In the first stage, we selected variables that were previously described in the literature that potentially contain predictive information. We then created Cox Proportional Hazards *single predictor models* (A) for each predictor and *interaction models* (B) which contain the predictor, treatment group assignment, and the interaction of predictor and treatment group assignment.

A. *Single predictor* Cox model results for each predictor with time to suicide event as the outcome. Missing values are imputed and pooled estimates were calculated using Rubin's rules. Predictors here are ordered by size of Optimism-Adjusted C. P-values less than 0.05 in bold.

<i>Predictor</i>	Hazard Ratio	Slope Estimate	Standard Error	Z Statistic	P-value	Observed C-statistic	Optimum Adjusted C-statistic
<i>Number of Hospitalizations</i>	1.27	0.24	0.08	2.85	<0.01	0.65	0.65
<i>Number of Suicide Attempts</i>	1.37	0.32	0.12	2.67	0.01	0.65	0.64
<i>Hamilton 17 Depression Inventory</i>	1.05	0.04	0.03	1.64	0.10	0.61	0.60
<i>Current Employment (Yes/No)</i>	3.31	1.20	0.53	2.25	0.02	0.60	0.60
<i>Childhood History of Abuse (Yes/No)</i>	0.47	-0.76	0.41	-1.88	0.06	0.60	0.59
<i>Reason for Living Scale (Total)</i>	0.99	-0.01	0.00	-1.81	0.07	0.60	0.59
<i>Beck Hopeless Scale (Total)</i>	1.06	0.05	0.03	1.61	0.11	0.59	0.59
<i>Age</i>	0.98	-0.02	0.02	-1.32	0.19	0.57	0.56
<i>Barrett Impulsivity Scale (Total)</i>	1.02	0.02	0.03	0.74	0.46	0.57	0.55
<i>Religion¹</i>	1.48	0.39	0.39	1.01	0.32	0.55	0.53
<i>Borderline personality disorder diagnosis (Yes/No)</i>	1.41	0.34	0.35	0.97	0.33	0.55	0.53
<i>Barrett Impulsivity Scale (Non-motor planning section)</i>	1.00	0.00	0.01	0.30	0.77	0.54	0.51
<i>Barrett Impulsivity Scale (Attentional)</i>	1.01	0.01	0.03	0.27	0.79	0.55	0.51
<i>Total definite suicide attempts in first degree relatives</i>	1.24	0.22	0.41	0.53	0.60	0.53	0.50
<i>Race²</i>	1.03	0.03	0.41	0.06	0.95	0.52	0.50
<i>Total definite suicide in first degree relatives</i>	0.89	-0.12	0.61	-0.20	0.84	0.50	0.48
<i>Alcohol Use Disorder (Yes/No)</i>	1.15	0.14	0.35	0.41	0.69	0.51	0.48

<i>Marital status</i> ³	1.00	0.00	0.41	0.01	0.99	0.51	0.48
<i>Substance Use Disorder other than alcohol (Yes/No)</i>	1.13	0.12	0.35	0.35	0.73	0.51	0.48
<i>Sex (Male/Female)</i>	0.95	-0.05	0.37	-0.14	0.89	0.51	0.48
<i>Any history of Substance Use Disorder including alcohol (Yes/No)</i>	0.94	-0.07	0.37	-0.18	0.86	0.51	0.47
<i>Treatment group assignment (Lithium vs. Valproate)</i>	0.93	-0.08	0.34	-0.22	0.82	0.51	0.47
<i>Ethnicity (Hispanic/non-Hispanic)</i>	0.99	-0.01	0.42	-0.03	0.98	0.50	0.47
<i>Tobacco smoking</i> ⁴	1.09	0.09	0.35	0.25	0.81	0.50	0.47
<i>Barrett Impulsivity Scale (Motor)</i>	0.99	-0.01	0.03	-0.34	0.73	0.50	0.47

B. Cox model results for interaction models. Using the same procedure for model fitting, we now also include predictor by treatment group interaction in the model. Table shows the predictor x treatment group estimates.

<i>Number of Hospitalizations</i>	1.02	0.02	0.17	0.09	0.93	0.65	0.62
<i>Number of Suicide Attempts</i>	0.77	-0.26	0.24	-1.08	0.28	0.65	0.62
<i>Current Employment (Yes/No)</i>	0.45	-0.81	1.06	-0.76	0.45	0.63	0.60
<i>Marital status</i> ³	0.10	-2.29	1.14	-2.00	0.05	0.63	0.59
<i>Hamilton 17 Depression Inventory</i>	0.94	-0.06	0.05	-1.09	0.28	0.62	0.58
<i>Reason for Living Scale (Total)</i>	1.00	-0.01	0.01	-0.58	0.57	0.61	0.56
<i>Childhood History of Abuse (Yes/No)</i>	1.36	0.31	0.82	0.38	0.71	0.60	0.56
<i>Beck Hopeless Scale (Total)</i>	0.98	-0.02	0.07	-0.33	0.74	0.60	0.56
<i>Age</i>	0.92	-0.08	0.04	-2.11	0.04	0.59	0.55
<i>Barratt Impulsivity Scale (Attentional)</i>	0.92	-0.08	0.05	-1.74	0.08	0.59	0.54
<i>Barratt Impulsivity Scale (Non-motor planning section)</i>	0.90	-0.10	0.08	-1.38	0.17	0.59	0.54
<i>Barratt Impulsivity Scale (Total)</i>	0.98	-0.03	0.02	-1.15	0.25	0.57	0.51
<i>Barratt Impulsivity Scale (Motor)</i>	1.00	0.00	0.05	0.07	0.95	0.57	0.50

<i>Religion</i> ¹	2.13	0.76	0.81	0.93	0.35	0.55	0.50
<i>Borderline personality disorder diagnosis (Yes/No)</i>	0.80	-0.22	0.71	-0.31	0.76	0.56	0.50
<i>Substance use disorder other than alcohol (Yes/No)</i>	2.47	0.90	0.71	1.28	0.20	0.55	0.50
<i>Race</i> ²	0.48	-0.74	0.84	-0.88	0.38	0.55	0.50
<i>Total definite suicide in first degree relatives</i>	0.17	-1.78	1.28	-1.39	0.17	0.55	0.50
<i>Tobacco smoking</i> ⁴	0.50	-0.69	0.71	-0.98	0.33	0.55	0.48
<i>Alcohol use disorder (Yes/No)</i>	1.80	0.59	0.71	0.83	0.40	0.54	0.47
<i>Total definite suicide attempts in first degree relatives</i>	0.56	-0.57	0.91	-0.63	0.53	0.54	0.47
<i>Ethnicity (Hispanic/non-Hispanic)</i>	0.96	-0.04	0.86	-0.05	0.96	0.52	0.46
<i>Any history of Substance Use Disorder including alcohol (Yes/No)</i>	0.89	-0.11	0.76	-0.15	0.88	0.51	0.44
<i>Sex (Male/Female)</i>	0.88	-0.12	0.73	-0.17	0.87	0.51	0.44

1. "How important is religion in your life?" 1- not at all important. 5 - extremely important

2. Self-Identified Racial group

- (1) "American Indian/Alaskan Native"
- (2) "Asian"
- (3) "Hawaiian/Pacific Islander"
- (4) "Black or African American"
- (5) "White"
- (6) "More than one race"
- (7) "Unknown or not reported"

3. Marital Status

- (1) "Single"
- (2) "Married"
- (3) "Separated"
- (4) "Divorced"
- (5) "Widowed"

4. Estimated daily consumption of tobacco

- (1) "Non-Smoker"
- (2) "Light"
- (3) "Moderate"
- (4) "Heavy"

Table 2. Summary Performance Measures of Predictive Models

	Predictor Terms Included in Model	Observed-C	Optimism-Adjusted C
Model 1 - Cox PH with main effects only			
	Number of Hospitalizations, Number of Suicide Attempts, Hamilton 17 Depression Inventory, Current Employment (Yes/No)	0.697	0.671
Model 2 - Cox PH model with predictors from Model 1 plus Treatment Group and Treatment by Predictor interactions			
	Number of Hospitalizations, Number of suicide attempts, Hamilton 17 Depression Inventory, Current Employment (Yes/No), Treatment Group, Number of Hospitalizations x Treatment Group, Number of suicide attempts x Treatment Group, Hamilton 17 Depression Inventory x Treatment Group, Current Employment (Yes/No) x Treatment Group	0.715	0.653
Model 3 - Cox PH model using LASSO with tuning parameters selected via 10-fold Cross Validation to minimize deviance			
	Age, Hamilton 17 Depression Inventory, Beck Hopeless Scale (Total), Number of Suicide Attempts, Childhood history of Abuse (Yes/No), Number of Hospitalizations, Current employment (Yes/No), and Reason for Living Scale (Total)	0.719	0.62
Model 4 - Cox PH model with all predictors and treatment group interactions, using LASSO with tuning			

parameters selected via 10-fold Cross Validation			
	Age, Hamilton 17 Depression Inventory, Beck Hopeless Scale (Total), Number of Suicide Attempts, Childhood History of Abuse (Yes/No), Number of Hospitalizations, Current employment (Yes/No), Reason for Living Scale (Total), Marital Status x Treatment interaction	0.733	0.622
Model 5 - Survival Tree			
	See Figure 1.	0.814	0.69

Table 3. Variable Importance for the top variables most frequently appearing in the Survival Tree models (larger number is more important), ranked by the order they appear for each imputed dataset from which they were trained (larger number bottom row more important for a particular imputation).

<i>Predictor</i>										
<i>Imputation Number</i>	numhosp	bistotn	bistot1	age	bistot2	ham17	rfltot	bistot3	curempl	numsui
1	15.49	10.81	5.59	12.82	7.51	3.22	8.54	7.8	2.53	3.71
2	21.93	11.11	5.09	10.67	7.12	3.08	8.64	8.15	7.96	1.89
3	12.59	9.58	14.14	5.3	8.49	9.31	9.49	2.78	3.82	0.76
4	19.5	7.4	11.93	11.62	9.74	13.4	2.48	4.04	6.5	0
5	11.97	18.62	9.82	7.57	12.41	7.88	5.27	5.53	4.24	6.11
<i>Mean Rank (larger most important)</i>	24.4	23	21	21	20.6	19.6	19.4	17.8	16.8	16.8

Abbreviations: Numhosp – number of hospitalizations; bistotn – Barrett Impulsivity Scale (total), bistot1 – Barrett Impulsivity Scale (attentional), age, bistot2 – Barrett Impulsivity Scale (motor), ham17 – Hamilton Depression Scale 17-item, rfltot – Reason for Living Scale, bistot3 – Barrett Impulsivity Scale (non-planning), curempl – current employment (Yes/No), numsui – number of previous suicide attempts.