New directions in machine learning analyses of administrative data to prevent suicide-related behaviors

Invited Commentary Amer J Epidemiology

Robert M. Bossarte
Chris J. Kennedy
Alex Luedtke
Matthew K. Nock
Jordan W. Smoller
Cara Stokes
Ronald C. Kessler

January 2021

Acknowledgements

Author Affiliations: Departments of Behavioral Medicine and Psychiatry, West Virginia University School of Medicine, Morgantown, West Virginia, United States; and U.S. Department of Veterans Affairs Center of Excellence for Suicide Prevention, Canandaigua, New York, United States (Robert M. Bossarte, Cara Stokes); Department of Biomedical Informatics, Harvard Medical School, Boston, MA, United States (Chris J. Kennedy); Department of Statistics, University of Washington, Seattle, Washington, United States; and Vaccine and Infectious Disease Division, Fred Hutchinson Cancer Research Center, Seattle, Washington, United States (Alex Luedtke); Department of Psychology, Harvard University, Cambridge, Massachusetts, United States (Matthew K. Nock); Department of Psychiatry, Massachusetts General Hospital, Boston, MA, United States (Jordan W. Smoller); Department of Health Care Policy, Harvard Medical School, Boston, Massachusetts, United States (Ronald C. Kessler).

Correspondence: Robert M Bossarte, Ph.D., Department of Behavioral Medicine and Psychiatry, West Virginia University School of Medicine, P.O. Box 9151, Morgantown, West Virginia, USA. Email: rbossarte@hsc.wvu.edu.

ORIGINAL UNEDITED MANUSCRIPT

© The Author(s) 2021. Published by Oxford University Press on behalf of the Johns Hopkins Bloomberg School of Public Health. All rights reserved. For permissions, please e-mail: journals.permissions@oup.com.

Downloaded from https://academic.oup.com/aje/advance-article/doi/10.1093/aje/kwab111/6239819 by Northwestern University Library user on 05 May 2021

Funding: This work was supported, in part, by the Department of Veterans Affairs Center of Excellence for Suicide Prevention (Bossarte) and the National Institute of Mental Health of the National Institutes of Health (R01MH121478, Kessler). The contents are solely the responsibility of the authors and do not necessarily represent the views of the funding organizations.

Conflict of interest: In the past 3 years, Dr. Kessler was a consultant for Datastat, Inc, Sage Pharmaceuticals, and Takeda. The other coauthors declare no conflicts.

Editor's note: *The opinions expressed in this article are those of the authors and do not necessarily reflect the views of the* American Journal of Epidemiology.

ABSTRACT

This issue contains a thoughtful report by Gradus et al. on a machine learning (ML) analysis of administrative variables to predict suicide attempts over two decades throughout Denmark. This is one of numerous recent studies that document strong concentration of risk of suicide-related behaviors (SRBs) among patients with high scores on ML models. The clear exposition of Gradus et al. provides an opportunity to review major challenges in developing, interpreting, and using such models: defining appropriate controls and time horizons, selecting comprehensive predictors, dealing with imbalanced outcomes, choosing classifiers, tuning hyperparameters, evaluating predictor variable importance, and evaluating operating characteristics. We close by calling for ML SRB research to move beyond merely demonstrating significant prediction, as this is by now well established, and to focus instead on using such models to target specific preventive interventions and to develop individualized treatment rules that can be used to help guide clinical decisions that address the growing problems of suicide attempts, suicide deaths, and other injuries and deaths in the same spectrum.

Keywords: suicide , prediction , machine learning

ORIGINAL UNEDITED MANUSCRIPT

Downloaded from https://academic.oup.com/aje/advance-article/doi/10.1093/aje/kwab111/6239819 by Northwestern University Library user on 05 May 2021

Abbreviations: AUC, area under the curve**;** CV, cross- validation; EHRs, electronic health records; ICD-10, International Classification of Diseases, Tenth Revision, Clinical Modification; ITR; individualized treatment rules; ML, machine learning; PPV, positive predictive value; PR, precision- recall; RCT, randomized controlled trial; RF; random forest; ROC; receiver operating characteristic; SAs, suicide attempts; SN, sensitivity; SP, specificity; SRBs, suicide related behaviors

Suicide remains a devastating and persistent global public health challenge. While advances in biomedicine have led to falling rates of many other leading causes of death, little progress has been made in reducing suicides. Indeed, the suicide rate has increased in many parts of the world over the past twenty years (1). Although our ability to predict suicide-related behaviors (SRBs) has been limited (2), efforts in recent years have leveraged large-scale data resources and powerful computational methods to overcome this problem. The report by Gradus and colleagues (3) is an excellent example of this line of research, using machine learning (ML) methods to predict administratively recorded nonfatal suicide attempts (SAs) throughout Denmark 1995-2015 from electronic health records (EHRs) and other administrative data. The report advances the field by leveraging nationwide registry data containing many predictors (n=1,458) and suicide attempts (n=22,974) to develop sex-specific ML models that achieve high accuracy (AUC=0.87-0.91) and, consistent with other recent studies (4-7), substantial concentrations of SAs in the high-risk tier (e.g., 40+% of incident SAs among the 5% of people with highest predicted risk). These results add to growing evidence that computational models from administrative data can stratify SRB risk at levels well beyond those achievable based exclusively on clinical evaluations.

At the same time, challenges exist in making methodologic choices in SRB prediction modeling (8,9). The clear and thoughtful exposition by Gradus et al. provides an opportunity to review best practices in these choices. We do that here, focusing on defining controls and risk

ORIGINAL UNEDITED MANUSCRIPT

Downloaded from https://academic.oup.com/aje/advance-article/doi/10.1093/aje/kwab111/6239819 by Northwestern University Library user on 05 May 2021

horizons, selecting predictors, dealing with imbalanced outcomes, choosing classifiers, tuning hyperparameters, evaluating predictor importance, and examining operating characteristics. We close by arguing that the time has come to begin developing clinically useful models tied to interventions and matching risk profiles with intervention options.

Defining cases and controls

Gradus et al. define cases as administratively recorded incident nonfatal SAs throughout Denmark 1995-2015. These are distinct from suicide deaths, which the authors studied in a separate paper (10). Suicide deaths are treated along with all other deaths in the current Gradus et al. paper as competing risks that result in patients being censored. It is noteworthy in this regard that previous research has shown that the predictors of suicide deaths are quite different from the predictors of nonfatal suicide attempts (11). For example, the suicide death rate is much higher among men than women whereas the nonfatal suicide attempt rate is much higher among women than men (12). Most of the methodological points we make below, however, apply equally to all SRBs whether they are fatal or nonfatal.

Gradus et al. define controls as randomly selected single months with no SA history as of the start date for a probability sample of "individuals living in Denmark on January 1, 1995." The inclusion of post-1994 immigrants as cases but not controls introduces bias, as Danish immigrants (8% of the population) have a much higher SA rate than non-immigrants (13, 14). This bias could have been removed either by requiring not only controls but also cases to be living in Denmark on January 1, 1995 or by removing this restriction from controls. The decision by Gradus et al. to select a random month for controls rather than sample by person-month proportional to cases in a case-crossover design with appropriate censoring could have

ORIGINAL UNEDITED MANUSCRIPT

Downloaded from https://academic.oup.com/aje/advance-article/doi/10.1093/aje/kwab111/6239819 by Northwestern University Library user on 05 May 2021

introduced additional biases. As reviewed elsewhere (8), similar issues in selecting appropriate controls are common in ML SRB studies.

Risk horizon

Gradus et al. create nested retrospective time intervals (from 0-6 to 0-48 months before SA) to aggregate EHR predictors involving visits, prescriptions, and procedures. This allows slopes to decay over longer lags (15). However, by including a 0-6 month retrospective time interval, the risk horizon (i.e., the survival period from last predictor assessment) is implicitly set at 0 months; that is, information from earlier in the same month as the SA is included among the predictors. This is not the optimal risk horizon for many applications, which would generally be prospective rather than cross-sectional. For example, a primary care provider seeing a patient for an annual physical might be concerned with SRB risk over the next 12 months, whereas an emergency department doctor deciding whether to hospitalize a suicidal patient might be more concerned with imminent risk (16). Only models with appropriate risk horizons can address these concerns. As important predictors change across risk horizons, some researchers develop separate models for different horizons using a fixed-point design rather than a retrospective case-control design. The retrospective creation of the predictor variables further presents a common form of temporal bias between the cases and the controls, which has recently been described (17). The differential time horizons of cases compared to controls has the effect of inflating the estimated discriminative performance relative to designs in which cases and controls have equivalent index events, such as the date of a medical visit or precursor diagnosis. Careful thought is consequently needed about the clinical decision the analysis is trying to mimic so as to make sure the temporal sampling in the design is appropriate for that purpose. Gradus et al. are unclear about this critical issue.

ORIGINAL UNEDITED MANUSCRIPT

Downloaded from https://academic.oup.com/aje/advance-article/doi/10.1093/aje/kwab111/6239819 by Northwestern University Library user on 05 May 2021

Predictors

Gradus et al. assembled an enviable set of predictors from high-quality Danish registries. Exploiting available predictor information in this way is critical for optimizing ML SRB models. Many investigators fail to do this. But some Key opportunities along these lines were not exploited by Gradus et al. Some examples:

It sometimes is useful to flip nested retrospective diagnostic codes across time intervals to learn about recency of first onset rather than most recent visit. For example, SRB risk associated with cancer is highest in the months after initial diagnosis and subsequently declines differentially by cancer type (18).

Along with aggregating 2-digit ICD-10 codes, it would have been useful to distinguish some disorder classes in a more fine-grained way given the existence of evidence about differences in SRB risk based on such distinctions (19). In addition, some cross-code composites, such as for chronic pain conditions (20) and the conjunction of primary and secondary site codes to identify metastasized cancer originating at a different site (21), predict SRBs. In predicting all SAs rather than incident SAs, information about prior SAs is critically important given that SA history is one of the strongest predictors of future SAs (22). Importantly, the predictors of incident SAs different from the predictors of repeat SAs (23).

Psychiatric treatment sector is also important, as psychiatric hospitalizations and emergency department visits are strong SRB predictors. In addition, ICD-10 S, T, and X codes capture information about external causes of injuries associated with exposure to violence, abuse and maltreatment that predict SRBs (24), whereas ICD-10 Z capture information about social determinants of health (e.g., homelessness, psychosocial needs, family stressors) that predict SRBs (25).

ORIGINAL UNEDITED MANUSCRIPT

Downloaded from https://academic.oup.com/aje/advance-article/doi/10.1093/aje/kwab111/6239819 by Northwestern University Library user on 05 May 2021

It is sometimes also possible to access and analyze clinical notes using natural language processing methods to elicit other information predicting SRBs (26).

Medication codes collapsed across therapeutic classes can be refined to distinguish medications within classes that protect against SRBs, such as lithium for bipolar disorder (27) and clozapine for schizophrenia (28) and to distinguish medications with suicide risk warnings or that pharmacoepidemiologic studies find predict SRBs (29).

Finally, in analyses spanning long time periods, as in the Gradus et al. study, it can be useful to include historical time as a predictor to investigate time trends in main effects and interactions. Gradus et al. found such an interaction involving sex that led to estimating sex-specific models. But other, perhaps more subtle, interactions with time could exist that could have been examined if time was included as a predictor and allowed to interact with other variables in the RF analysis.

Imbalanced data

Most algorithms produce suboptimal results when predicting rare dichotomous outcomes unless adjustments are made (30). Numerous methods developed for this purpose differentially weight or sample cases versus controls (data-based methods) or, like Gradus et al., weight the algorithm loss function to penalize false negatives more than false positives (classifier-based methods; 31). As model performance can vary based on this decision, it is useful to explore a range of penalties (32).

Choosing classifiers

No single ML classifier is optimal for all prediction problems. Kaggle competitions typically find that Random Forest (RF; the classifier Gradus et al. used), various kinds of gradient boosting, (XGBoost, Lightgbm, Catboost, or a stacked emsemble of more than one of

ORIGINAL UNEDITED MANUSCRIPT

Downloaded from https://academic.oup.com/aje/advance-article/doi/10.1093/aje/kwab111/6239819 by Northwestern University Library user on 05 May 2021

these options) and BART outperform other methods in predicting structured data like in ML SRB models (33, 34). Relative algorithm performance varies across applications, though, and promising new algorithms are constantly emerging (35). Some researchers address this by replicating analyses with several different algorithms in a training sample before picking the best classifier for their final model. As noted by Gradus et al., though, this is unnecessary, as an ensemble method exists that uses cross-validation (CV) to create an optimal weighted combination of predicted outcome scores across multiple classifiers that is guaranteed in expectation to perform at least as well as the best component algorithm (36, 37).

Tuning hyperparameters

Prediction models typically require some model characteristics to be fixed (hyperparameters) before estimation. Changing (tuning) these hyperparameters often leads to substantial prediction improvement (38, 39). Gradus et al. followed standard practice for fixed RF hyperparameter values, but a data-driven method to optimize hyperparameter selection should be used when development is for implementation rather than demonstration (40).

Evaluating predictor importance

Gradus et al. focus on predictor variable importance. This is understandable, as researchers and clinicians are interested in the predictors that drive model performance. It is critical to recognize, though, that ML models prioritize prediction accuracy over variable interpretation, resulting in ML importance metrics having limited value. For example, the several different importance metrics in RF typically produce distinct variable importance rankings, all of which are: (i) biased by favoring predictors with many values and low correlations with other predictors; and (ii) internally inconsistent (i.e., estimated variable importance can go down when the model is changed to rely more on that variable; 41, 42).

ORIGINAL UNEDITED MANUSCRIPT

Downloaded from https://academic.oup.com/aje/advance-article/doi/10.1093/aje/kwab111/6239819 by Northwestern University Library user on 05 May 2021

A recently developed procedure called *TreeExplainer* resolves some of these problems (43) and introduces procedures to discover critical interactions and differences in RF variable importance across individuals (44). Nonetheless, caution is still needed in making substantive interpretations of such results. Prediction importance should not be confused with causal importance. Instead, predictor importance analysis is most useful for pruning predictors to increase out-of-sample model performance. If researchers have a serious interest in investigating variable importance for targeting interventions, a better approach is to assign a predicted probability of SRB to each observation based on the ML model, apply precision-recall curves or net benefit curves (see next section) to select an intervention decision threshold, and inspect clustering of modifiable predictors that distinguish above versus below that threshold for targeted learning analyses (45) to evaluate potential implications of preventive interventions.

Operating characteristics

The Gradus et al. RF models have excellent operating characteristics. For example, 44.7% of incident male SAs occurred among the 5% of males with highest CV risk. However, given the rarity of SA (87/100,000 male person-years), precision-recall (PR) curves should also be estimated to examine associations between positive predictive value ("precision") and sensitivity ("recall"). This would allow us to see, for example, that 6-month PPV is only about 0.4% among the 5% of men with highest predicted SA risk, a fact that was not emphasized by Gradus et al. Just as the AUC-ROC summarizes the association between SN and 1-SP in an ROC curve, the AUC-PR summarizes the association between PPV and SN in a PR curve. This becomes especially important when ML is used to guide clinical decision-making, as joint understanding of risk above a decision threshold in the absence of intervention and likely intervention effectiveness are needed to convey information about clinical implications of an

ORIGINAL UNEDITED MANUSCRIPT

Downloaded from https://academic.oup.com/aje/advance-article/doi/10.1093/aje/kwab111/6239819 by Northwestern University Library user on 05 May 2021

intervention (46). When intervention costs and benefits are known, it is also useful to calculate a net benefit curve to arrive at a principled decision threshold (47). As a best practice, it is also useful to examine the calibration of the final model – how similar the model's probability predictions are to the observed probabilities of SA (48). Replicability and transparency can be enhanced further by following model reporting guidelines (e.g., 49).

Where should we go from here?

In addition to the above suggestions for optimizing prediction and interpretation, it is noteworthy that other types of data are increasingly being used to improve ML SRB models, (e.g., patient-report surveys, digital devices, biomarkers; 50-52). As these can be burdensome and expensive, though, careful consideration is needed of incremental benefits. A tiered approach can do this by beginning with passively-collected administrative data like those used by Gradus et al., with a focus on ruling out low-risk patients from further assessments (8). Subsequent steps can then include inexpensive self-report surveys followed by additional ML analyses to target successively more refined subsets of patients for more intensive-expensive asessments culminating in in-depth clinical evaluations based on all data (9).

Future ML SRB research also needs to be more focused on why ML SRBs are being developed. Gradus et al. note that such models have two purposes: to provide insights into important SRB predictors; and to pinpoint high-risk patients for targeted interventions. We already noted the limitations of ML importance measures in achieving the first purpose. Two issues can also be raised about the second purpose. One was already noted: that models designed to target preventive interventions should select samples and time horizons appropriate to those interventions (53). This is not always done in existing aplications (9,17).

The other is that intervention assignment should be based ideally on an understanding of

ORIGINAL UNEDITED MANUSCRIPT

Downloaded from https://academic.oup.com/aje/advance-article/doi/10.1093/aje/kwab111/6239819 by Northwestern University Library user on 05 May 2021

not only SRB risk but also comparative effects of intervention options for specific patients. For example, controlled trials show that several outpatient therapies reduce SRBs among patients reporting suicidality (54). Conventional ML SRB prediction models might do an excellent job determining which patients are most in need of these interventions. However, conventional ML SRB prediction models do not help determine which intervention is best for which patient. Individualized treatment rules (ITRs) providing this information would be of great value.

ML models to develop ITRs are different from models to predict SRB risk because patients at highest risk are not necessarily those most likely to be helped by available interventions. ITR models instead evaluate interactions between presumed *prescriptive* predictors (i.e., predictors of greater response to one intervention than others) and intervention type received (55). Although ITR models should ideally be developed in comparative effectiveness trials, such trials can sometimes be emulated in observational samples (56). Such adjusted databases can yield results close to those of RCTs (57). Extensions exist to develop ITRs that estimate the subset of patients that would benefit from intervention (58).

Conclusions

We do not want to leave the impression that the Gradus et al. study is an inferior example of an ML SRB study. It is not. The problems we described exist throughout the ML SRB literature and more generally in epidemiologic studies focused narrowly on prediction rather than broader considerations of impact or use (59). It is by now well-known that ML methods can be used to predict SRBs. We consequently need to take the next step of applying thoughtful versions of ML SRB models to help target and evaluate carefuly selected interventions in appropriate samples and with risk horizens appropriate to these interventions. This should be followed by subsequent ML analyses to develop ITRs for comparative effectiveness of

ORIGINAL UNEDITED MANUSCRIPT

Downloaded from https://academic.oup.com/aje/advance-article/doi/10.1093/aje/kwab111/6239819 by Northwestern University Library user on 05 May 2021

alternative interventions aimed at optimizing interventions to address the growing problems of SAs, suicides, and other external causes of injury and death in the same spectrum (60).

REFERENCES

1. Naghavi M. Global, regional, and national burden of suicide mortality 1990 to 2016: systematic analysis for the Global Burden of Disease Study 2016. *BMJ*. 2019;364:l94.
2. Franklin JC, Ribeiro JD, Fox KR, et al. Risk factors for suicidal thoughts and behaviors: A meta-analysis of 50 years of research. *Psychol Bull*. 2017;143(2):187–232.
3. Gradus JL, Rosellini AJ, Horváth-Puhó E, et al. Predicting sex-specific non-fatal suicide attempt risk using machine learning and data from Danish national registries. *Am J Epidemiol*. This issue.
4. Belsher BE, Smolenski DJ, Pruitt LD, et al. Prediction models for suicide attempts and deaths: A systematic review and simulation. *JAMA Psychiatry*. 2019;76(6):642–651.
5. Burke TA, Ammerman BA, Jacobucci R. The use of machine learning in the study of suicidal and non-suicidal self-injurious thoughts and behaviors: a systematic review. *J Affect Disord*. 2019;245:869–884.
6. Bernert RA, Hilberg AM, Melia R,et al. Artificial intelligence and suicide prevention: A systematic review of machine learning investigations. *Int J Environ Res Public Health.* 2020;17(16):5929.
7. McHugh CM, Large MM. Can machine-learning methods really help predict suicide? *Curr Opin Psychiatry*. 2020;33(4):369–374.

ORIGINAL UNEDITED MANUSCRIPT

Downloaded from https://academic.oup.com/aje/advance-article/doi/10.1093/aje/kwab111/6239819 by Northwestern University Library user on 05 May 2021

8. Kessler RC, Bernecker SL, Bossarte RM, et al. The role of big data analytics in predicting suicide. In: Passos I, Mwangi B, Kapczinski F (eds) *Personalized Psychiatry*: Springer, Cham; 2019:77–98.

9. Kessler RC, Bossarte RM, Luedtke A,et al. Suicide prediction models: A critical review of recent research with recommendations for the way forward. *Mol Psychiatry*. 2020;25(1):168–179.

10. Gradus JL, Rosellini AJ, Horváth-Puhó E, et al. Prediction of sex-specific suicide risk using machine learning and single-payer health care registry data from Denmark. *JAMA Psychiatry*. 2020;77(1):25–34.

11. Klonsky ED, May AM, Saffer BY. Suicide, suicide attempts, and suicidal ideation. *Annu Rev Clin Psychol.* 2016;12:307–330.

12. Fox KR, Millner AJ, Mukerji CE, et al. Examining the role of sex in self-injurious thoughts and behaviors. *Clin Psychol Rev*. 2018;66:3–11.

13. Denmark Statistics. Immigrants and their descendants. https://www.dst.dk/en/Statistik/emner/befolkning-og-valg/indvandrere-og-efterkommere. Accessed March 15, 2021.

14. Webb RT, Antonsen S, Pedersen CB, et al. Attempted suicide and violent criminality among Danish second-generation immigrants according to parental place of origin. *Int J Soc Psychiatry*. 2016;62(2):186–197.

15. Barak-Corren Y, Castro VM, Javitt S, et al. Predicting suicidal behavior from longitudinal electronic health records. *Am J Psychiatry*. 2017;174(2):154–162.

16. Galynker I, Galynker II. *The suicidal crisis: clinical guide to the assessment of imminent suicide risk*. New York, NY: Oxford University Press; 2017.

ORIGINAL UNEDITED MANUSCRIPT

Downloaded from https://academic.oup.com/aje/advance-article/doi/10.1093/aje/kwab111/6239819 by Northwestern University Library user on 05 May 2021

17. Yuan W, Beaulieu-Jones BK, Yu KH, et al. Temporal bias in case-control design: preventing reliable predictions of the future. *Nat Commun.* 2021;12(1):1107.

18. Zaorsky NG, Zhang Y, Tuanquin L,et al. Suicide among cancer patients. *Nat Commun.* 2019;10(1):207.

19. Urban D, Rao A, Bressel M,et al. Suicide in lung cancer: who is at risk? *Chest*. 2013;144(4):1245–1252.

20. Ilgen MA, Kleinberg F, Ignacio RV, et al. Noncancer pain conditions and risk of suicide. *JAMA Psychiatry*. 2013;70(7):692–697.

21. Rahouma M, Kamel M, Abouarab A, et al. Lung cancer patients have the highest malignancy-associated suicide rate in USA: a population-based analysis. *Ecancermedicalscience.* 2018;12:859.

22. Borges G, Angst J, Nock MK, et al. A risk index for 12-month suicide attempts in the National Comorbidity Survey Replication (NCS-R). *Psychol Med*. 2006;36(12):1747–1757.

23. Pagura J, Cox BJ, Sareen J, et al. Factors associated with multiple versus single episode suicide attempts in the 1990-1992 and 2001-2003 United States national comorbidity surveys. *J Nerv Ment Dis.* 2008;196(11):806–813.

24. Haglund A, Lindh Å U, Lysell H, et al. Interpersonal violence and the prediction of short-term risk of repeat suicide attempt. *Sci Rep*. 2016;6:36892.

25. Blosnich JR, Montgomery AE, Dichter ME, et al. Social determinants and military veterans' suicide ideation and attempt: A cross-sectional analysis of electronic health record data. *J Gen Intern Med*. 2020;35(6):1759–1767.

ORIGINAL UNEDITED MANUSCRIPT

Downloaded from https://academic.oup.com/aje/advance-article/doi/10.1093/aje/kwab111/6239819 by Northwestern University Library user on 05 May 2021

26. Levis M, Leonard Westgate C, Gui J, et al. Natural language processing of clinical mental health notes may add predictive value to existing suicide risk models. *Psychol Med*. 2020:1–10.

27. Caley CF, Perriello E, Golden J. Antiepileptic drugs and suicide-related outcomes in bipolar disorder: A descriptive review of published data. *Ment Health Clin*. 2018;8(3):138–147.

28. Taipale H, Lähteenvuo M, Tanskanen A, et al. Comparative effectiveness of antipsychotics for risk of attempted or completed suicide among persons with schizophrenia. *Schizophr Bull.* 2020; sbaa111.

29. Gibbons R, Hur K, Lavigne J, et al. Medications and suicide: High Dimensional Empirical Bayes Screening (iDEAS). *Harvard Data Science Review*. 2019;1(2).

30. López V, Fernández A, García S, et al. An insight into classification with imbalanced data: Empirical results and current trends on using data intrinsic characteristics. *Inf Sci*. 2013;250:113–141.

31. Haixiang G, Yijing L, Shang J, et al. Learning from class-imbalanced data: Review of methods and applications. *Expert Syst Appl*. 2017;73:220–239.

32. Leevy JL, Khoshgoftaar TM, Bauder RA, et al. A survey on addressing high-class imbalance in big data. *J Big Data*. 2018;5(1):42.

33. Harasymiv V. Lessons from 2 million machine learning models on Kaggle. KDnuggets 15:n42. https://www.kdnuggets.com/2015/12/harasymiv-lessons-kaggle-machine-learning.html. Accessed December 27, 2020.

ORIGINAL UNEDITED MANUSCRIPT

Downloaded from https://academic.oup.com/aje/advance-article/doi/10.1093/aje/kwab111/6239819 by Northwestern University Library user on 05 May 2021

34. Fogg A. Anthony goldbloom gives you the secret to winning kaggle competitions. https://www.import.io/post/how-to-win-a-kaggle-competition/. Published January 13, 2016. Accessed December 27, 2020.

35. Hancock JT, Khoshgoftaar TM. CatBoost for big data: An interdisciplinary review. *J Big Data*. 2020;7(1):94.

36. Van der Laan MJ, Polley EC, Hubbard AE. Super learner. Stat Appl Genet Mol Biol. 2007;6:Article 25.

37. Kennedy C. Guide to SuperLearner. https://cran.r-project.org/web/packages/SuperLearner/vignettes/Guide-to-SuperLearner.html. Published March 16, 2017. Accessed December 27, 2020.

38. Thornton C, Hutter F, Hoos HH, et al. Auto-weka: Automated selection and hyper-parameter optimization of classification algorithms. *CoRR*, abs/1208.3719. 2012.

39. Koch P, Golovidov O, Gardner S, et al. Autotune: a derivative-free optimization framework for hyperparameter tuning. In: *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. 2018; pp. 443- 452).

40. Probst P, Wright MN, Boulesteix AL. Hyperparameters and tuning strategies for random forest. *Wiley Interdiscip Rev Data Min Knowl Discov*. 2019;9(3):e1301.

41. Strobl, C., Boulesteix, AL., Zeileis, A. et al. Bias in random forest variable importance measures: Illustrations, sources and a solution. BMC Bioinformatics 8, 25 (2007). https://doi.org/10.1186/1471-2105-8-25

42. Lundberg S. Interpretable machine learning with XGBoost. Towards Data Science. https://towardsdatascience.com/interpretable-machine-learning-with-xgboost-9ec80d148d27. Published 17, 2018. Accessed December 16, 2020.

ORIGINAL UNEDITED MANUSCRIPT

Downloaded from https://academic.oup.com/aje/advance-article/doi/10.1093/aje/kwab111/6239819 by Northwestern University Library user on 05 May 2021

43. Lundberg SM, Erion G, Chen H, et al. From local explanations to global understanding with explainable AI for trees. *Nat Mach Intell.* 2020;2(1):56–67.

44. Lundberg SM, Lee S-I. A unified approach to interpreting model predictions. In: *NIPS'17 Proceedings of the 31st International Conference on Neural Information Processing Systems*, vol 30, 2017; pp. 4765–4774.

45. Van der Laan MJ, Rose S. *Targeted learning in data science*. New York, New York: Springer International Publishing; 2018.

46. Saver JL, Lewis RJ. Number needed to treat: Conveying the likelihood of a therapeutic effect. *JAMA*. 2019;321(8):798–799.

47. Vickers AJ, Van Calster B, Steyerberg EW. Net benefit approaches to the evaluation of prediction models, molecular markers, and diagnostic tests. *BMJ*. 2016;352:i6.

48. Austin PC, Steyerberg EW. The Integrated Calibration Index (ICI) and related metrics for quantifying the calibration of logistic regression models. *Stat Med.* 2019;38(21):4051–4065.

49. Collins GS, Reitsma JB, Altman DG, et al. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): the TRIPOD statement. The TRIPOD Group. *Circulation.* 2015;131(2):211–219.

50. Glenn CR, Nock MK. Improving the short-term prediction of suicidal behavior. *Am J Prev Med*. 2014;47(3 Suppl 2):S176–180.

51. Vahabzadeh A, Sahin N, Kalali A. Digital suicide prevention: Can technology become a game-changer? *Innov Clin Neurosci*. 2016;13(5-6):16–20.

52. Sudol K, Mann JJ. Biomarkers of suicide attempt behavior: Towards a biological model of risk. *Curr Psychiatry Rep*. 2017;19(6):31.

ORIGINAL UNEDITED MANUSCRIPT

Downloaded from https://academic.oup.com/aje/advance-article/doi/10.1093/aje/kwab111/6239819 by Northwestern University Library user on 05 May 2021

53. Kessler RC, Bauer MS, Bishop TM, et al. Using administrative data to predict suicide after psychiatric hospitalization in the Veterans Health Administration system. *Front Psychiatry*. 2020;11:390.

54. Jobes DA, Au JS, Siegelman A. Psychological approaches to suicide treatment and prevention. *Curr Treat Options Psychiatry*. 2015;2(4):363–370.

55. L Luedtke AR, van der Laan MJ. Statistical inference for the mean outcome under a possibly non- unique optimal treatment strategy. *Ann Stat*. 2016;44(2):713–742.

56. Hernán MA, Robins JM. Using big data to emulate a target trial when a randomized trial is not available. *Am J Epidemiol*. 2016;183(8):758–764.

57. Anglemyer A, Horvath HT, Bero L. Healthcare outcomes assessed with observational study designs compared with those assessed in randomized trials. *Cochrane Database Syst Rev*. 2014(4):Mr000034.

58. Luedtke AR, van der Laan MJ. Evaluating the impact of treating the optimal subgroup. *Stat Methods Med Res*. 2017;26(4):1630–1640.

59. Galea S. An argument for a consequentialist epidemiology. *Am J Epidemiol*. 2013;178(8):1185–1191.

60. Case A, Deaton A. *Deaths of despair and the future of capitalism*. Princeton, New Jersey: Princeton University Press; 2020.

ORIGINAL UNEDITED MANUSCRIPT

Downloaded from https://academic.oup.com/aje/advance-article/doi/10.1093/aje/kwab111/6239819 by Northwestern University Library user on 05 May 2021