Psychiatry Research 323 (2023) 115175

Contents lists available at ScienceDirect

Psychiatry Research

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

An efficient landmark model for prediction of suicide attempts in multiple clinical settings

Yi-han Sheu [a,b,c,d,1], Jiehuan Sun [e,1], Hyunjoon Lee [a,b,c], Victor M. Castro [c,f], Yuval Barak-Corren [g,h], Eugene Song [a,b,c], Emily M. Madsen [a,b,c], William J. Gordon [i,j], Isaac S. Kohane [i,j], Susanne E. Churchill [i,j], Ben Y. Reis [g,i], Tianxi Cai [i,k,l,1], Jordan W. Smoller [a,b,c,d,1,*]

[a] Center for Precision Psychiatry, Massachusetts General Hospital, 185 Cambridge Street, Boston, MA 02114, USA
[b] Psychiatric and Neurodevelopmental Genetics Unit, Massachusetts General Hospital, 185 Cambridge Street, Boston, MA 02114, USA
[c] Department of Psychiatry, Harvard Medical School, 401 Park Drive, Boston, MA 02215, USA
[d] Broad Institute of MIT and Harvard, 415 Main St, Cambridge, MA 02142, USA
[e] Department of Epidemiology and Biostatistics, University of Illinois Chicago, 1603W. Taylor St., Chicago, IL 60612, USA
[f] Center for Quantitative Health, Massachusetts General Hospital, 185 Cambridge Street, Boston, MA 02114, USA
[g] Department of Pediatrics, Boston Children's Hospital, 300 Longwood Ave, Boston, MA 02115, USA
[h] Schneider Children's Medical Center of Israel, 14 Kaplan Street, Petaḥ Tiqwa, Central, Israel
[i] Department of Biomedical Informatics, Harvard Medical School, 10 Shattuck Street, Boston, MA 02115, USA
[j] Department of Medicine, Brigham and Women's Hospital, 75 Francis Street, Boston, MA 02115, USA
[k] Department of Biostatistics, Harvard T.H. Chan School of Public Health, 677 Huntington Avenue, Boston, MA 02115, USA
[l] Translational Data Science Center for a Learning Health System, Harvard University, 677 Huntington Avenue, Boston, MA, USA

ARTICLE INFO

Keywords:
Suicide attempt
Prediction
Electronic health record
Landmark model

ABSTRACT

Growing evidence has shown that applying machine learning models to large clinical data sources may exceed clinician performance in suicide risk stratification. However, many existing prediction models either suffer from "temporal bias" (a bias that stems from using case-control sampling) or require training on all available patient visit data. Here, we adopt a "landmark model" framework that aligns with clinical practice for prediction of suicide-related behaviors (SRBs) using a large electronic health record database. Using the landmark approach, we developed models for SRB prediction (regularized Cox regression and random survival forest) that establish a time-point (e.g., clinical visit) from which predictions are made over user-specified prediction windows using historical information up to that point. We applied this approach to cohorts from three clinical settings: general outpatient, psychiatric emergency department, and psychiatric inpatients, for varying prediction windows and lengths of historical data. Models achieved high discriminative performance (area under the Receiver Operating Characteristic curve 0.74–0.93 for the Cox model) across different prediction windows and settings, even with relatively short periods of historical data. In short, we developed accurate, dynamic SRB risk prediction models with the landmark approach that reduce bias and enhance the reliability and portability of suicide risk prediction models.

1. Introduction

Suicide rates have increased by more than 30% over the past two decades (National Center for Health Statistics, 2018) and death by suicide is the second leading cause of death among 10–34 year-olds (National Center for Health Statistics, 2018). Most individuals who attempt or die by suicide are seen by a healthcare provider in the preceding month (Ahmedani et al., 2014), making healthcare settings a key venue for identifying and intervening to prevent suicide-related behavior (SRB). Unfortunately, studies have shown that clinician performance in predicting suicide risk is little better than chance (Nock et al., 2022).

Recently, however, a growing number of studies have demonstrated

* Corresponding author at: Center for Precision Psychiatry, Massachusetts General Hospital, 185 Cambridge Street, Boston, MA 02114, USA.
E-mail address: jsmoller@mgh.harvard.edu (J.W. Smoller).
[1] These authors equally contributed to this work.

https://doi.org/10.1016/j.psychres.2023.115175
Received 20 November 2022; Received in revised form 16 March 2023; Accepted 18 March 2023
Available online 21 March 2023

that applying statistical and machine learning models to the vast clinical data sources found in health system records can enable risk stratification that appears to exceed clinician performance (Barak-Corren et al., 2017, 2020; Chen et al., 2020; Cho et al., 2020; Fisher et al., 2022; Karmakar et al., 2016; Levis et al., 2020, 2022; Machado et al., 2022; Malone et al., 2021; Nock et al., 2022; Nordin et al., 2021; Obeid et al., 2020; Rossom et al., 2021; Simon et al., 2018, 2019; Su et al., 2020; Tsui et al., 2021; van Mens et al., 2020; Walsh et al., 2017, 2018; Wei et al., 2021; Xu et al., 2022; Zheng et al., 2020). We have recently shown that implementation of these models can be cost-effective in reducing SRBs when paired with evidence-based interventions (Ross et al., 2021). However, existing prediction models have had key limitations. Many models reported to date (Barak-Corren et al., 2017; Cohen et al., 2020; Levis et al., 2020; Nordin et al., 2021; Obeid et al., 2020; Tsui et al., 2021; van Mens et al., 2020; Walsh et al., 2018; Xu et al., 2022) have been trained using a case-control approach in which predictors are selected and weighted according to their differential frequency between individuals who engaged in SRBs (cases) and those who did not (controls). As such, the trajectory of predictive features for cases are collected up to the point of the SRB, while those sampled for controls occur throughout the period of observation available for these individuals. This leads to a fundamental mismatch in the distribution of data between the inputs for model development, and what clinicians see during actual practice (Yuan et al., 2021). This "temporal bias" (Yuan et al., 2021) makes case-control model training quite different from the real-world scenario in which a clinician would utilize the model for prospective prediction. In practice, a clinician might wish to know whether a given individual is likely to attempt suicide within a specific window of time (e.g. the next month). However, the clinician would not have access to the (future) case vs. control status of a given individual – the outcome on which the model was trained. As Yuan and colleagues have shown (Yuan et al., 2021), this bias may result in a spurious inflation of effect sizes and render the model unsuitable for prospective prediction.

Prior models that adopt a visit-based training scheme (i.e., using healthcare visits, instead of patients, as units of instances for model development) (Chen et al., 2020; Cho et al., 2020; Karmakar et al., 2016; Simon et al., 2018; Zheng et al., 2020) are typically not affected by temporal bias induced by case-control sampling because they are based on prospective designs and were trained irrespective of future case status. However, previous modeling schemes have either been (1) trained using all available visits (Chen et al., 2020; Simon et al., 2018), which potentially biased towards patients with more healthcare utilizations; or (2) limited to specific time windows or settings (Cho et al., 2020; Karmakar et al., 2016; Zheng et al., 2020)

A major goal of developing and validating suicide prediction algorithms is to enable the development of clinical decision support tools that could be deployed at the point-of-care with scalability, flexibility, and transportability. However, as noted above, most attempts to construct such algorithms are either susceptible to temporal bias or bias towards patients with higher healthcare utilization that could compromise their performance, or have computational requirements that may limit their use in resource-constrained settings. To address these problems, we apply a systematic approach to prediction based on a "landmark model" framework (Parast and Cai, 2013; Parast et al., 2012; Van Houwelingen, 2007) which ameliorates the aforementioned issues through proper sampling. The main advantage of this approach is that by design, models can be flexibly trained in a manner that mirrors the clinical situation that practitioners actually face (in which future outcomes are not known); this avoids biases arising from distributional misalignments, promoting model interpretation and fairness. In essence, landmark modeling is based on defining a time point of interest (the "landmark time" – which can be one or more points of time of any type whenever data allows – e.g., at a certain type of clinical visit, or upon receiving a certain kind of diagnosis), from which prospective prediction is made without pre-specifying subsequent case or control status. Data are then sampled for all patients prior to the landmark time(s) and predictions are made for pre-defined prediction windows (e.g. the next month). This ensures that the distributions of patient training data (1) are not systematically different for those who later do or do not engage in SRBs; and (2) can be aligned with the particular application settings of interest. This approach can be used with whatever statistical or machine learning model architecture desired.

Here, we apply landmark modeling to data from longitudinal electronic health record (EHR) data from a large healthcare system (Mass General Brigham (MGB)) and train suicide risk prediction models using two approaches: one sparse linear model (regularized Cox regression (Wu, 2012)) and one model that accommodates complex interactions (random survival forest (Ishwaran et al., 2008)).

2. Methods

2.1. Study populations and cohort definition

The data for the study were extracted from the MGB Research Patient Data Registry (RPDR) (Nalichowski et al., 2006). The RPDR is a centralized data registry that gathers clinical information from the MGB EHR. The RPDR database includes more than 7 million patients with over 3 billion records seen across more than 8 hospitals, including two major teaching hospitals: Massachusetts General Hospital and Brigham and Women's Hospital. We defined three patient cohorts of interest within the RDPR by clinical settings: (1) a "general outpatient cohort", which includes all patients who had at least 3 visits and a minimum of 90 days of medical record, and were aged between 15 and 85 with at least one outpatient visit during the period of Jan 1, 2016 and Dec. 31, 2018 ("the study time frame"). The starting date was chosen to reflect the completion of conversion from ICD-9 to ICD-10 to reduce heterogeneity in feature documentation; (2) a "psychiatric ED cohort", applying the same age and data requirements as the general outpatient cohort, but requiring at least one emergency department visit (instead of an outpatient visit) within the study time frame, during which there was psychiatric evaluation/consultation; and (3) a "psychiatric inpatient cohort", applying the same age and data requirements as the previous two cohorts, but instead requiring at least one psychiatric inpatient admission during the study time frame. For each of these cohorts, we randomly sampled one visit per patient (i.e., any outpatient visit for the general outpatient cohort; a psychiatric-ED visit for the psychiatric ED cohort, and a psychiatric inpatient admission for the psychiatric inpatient cohort) during the study timeframe and used it as the "landmark visit" (i.e., the visit from which a prediction is made). For psychiatric-ED and inpatient landmark visits, where encounters may span more than one day, the last day of the visit was the time when prediction was performed.

2.2. Suicide-related behavior definition

SRB was defined using a set of International Classification of Disease 10 Clinical Modification (ICD-10-CM (World Health Organization, 1993)) codes as previously reported (Barak-Corren et al., 2020, 2017). As previously described (Barak-Corren et al., 2017), these codes were shown to be valid for capturing intentional self-harm by extensive chart review by expert clinicians, with a positive predictive value (PPV) greater than 70%. To further increase the PPV of the SRB ICD codes for outcome measurement, we limited the codes to those which are recorded at either an emergency department (ED) or a psychiatric inpatient encounter for outcome events. As such, non-ED or psychiatric inpatient SRB ICD codes were included as prediction features, but not as outcome SRB events.

2.3. Feature extraction and outcome definition

The features and outcomes were defined in the same way for all the three cohorts defined above (i.e., general outpatient, ED, and psychiatric

inpatient). Specifically, the survival outcome was defined as the time to the first SRB-related diagnostic code since the landmark visit. SRB codes at the first day after the landmark visit were not included to ensure SRB indeed occurred after the landmark visit, as ICD code entries can be delayed. In addition, SRB codes that occur within five days after the last SRB code prior to or on the last day of the landmark visit are not included to minimize the probability that the SRB code represents an already documented, rather than a new SRB event.

The EHR prediction features were derived using historical EHR data (ICD codes) prior to or at the landmark visit, as follows. We defined several "look back" periods (0.5, 1, 2, 4, 6, 8 years and all historical data). representing the time frame during which historical features are included. Feature sets were constructed separately for each look back period. For example, EHR features included in the two-year look back period were only those documented on or prior to the landmark visit for all patients having more than two years' data. In this case, we used all EHR data on or prior to the landmark visit for patients having less than two years' data. We then converted historical ICD codes to PheWAS codes ("phecodes"), derived from the PheWAS catalog (https://phewascatalog.org/phecodes) that groups related ICD codes into clinically relevant groups (Denny et al., 2013). Next, the total count of each phecode during the look back period was extracted for each patient (1799 phecodes in total). Duplicated phecodes on the same day are only counted once to avoid "double counting". We also removed rare phecodes (occurring in fewer than 10 patients in each of the study cohorts). To take into account the different lengths of history available in our database, we then normalized the total phecode count to the length of observation time, followed by log-transformation. Thus, for the two-year look back period, phecode_normalized = log(phecode_count + 1) – log (T_obs), where T_obs is equal to two for patients having more than two years' data and is equal to the number of years between the first visit (defined by the first ICD code ever recorded in the database) and the landmark visit for patients having less than two years' data. We also include the number of prior SRBs (as defined above) as a feature (after normalization and log transformation).

In addition to features based on phecodes, we derived a feature ("HU_normalized") indexing overall healthcare utilization by extracting the total count of unique visits in the look back period and then normalizing the count to the length of observation time for the data, followed by log-transformation. Lastly, we included demographics features (i.e., gender, race, age, and public payer status). Prior to model training, each feature is standardized to have a mean of zero and a standard deviation of one.

2.4. Model building and evaluation

We evaluated two models for risk prediction: regularized Cox regression (i.e., regularized Cox models with linear effects for survival outcomes) and random survival forest (RF, the survival version of a non-linear tree-based model). We constructed the prediction models for each look back period and each setting separately. To improve the computational efficiency for the general outpatient cohort, where there are a large number of visits with no subsequent SRB events in the prediction windows ("non-event visits,"), we trained models using a fraction (1/20) of randomly sampled non-event visits, and used all visits followed by at least one SRB event. The effect of down-sampling was then reverted to the full EHR data size by up-weighting non-event visits during model training by a factor of 20. We used the R packages *glmnet* (Simon et al., 2011) and *randomForestSRC* (Ishwaran et al., 2008) to fit the regularized Cox models and random survival forests, respectively.

For the regularized Cox model, we used nested cross-validation (with K=5 folds for both inner and outer splits) for hyperparameter tuning (via inner splits) and evaluation the predictive performance of the models (via outer splits). The models were trained utilizing both L1 and L2 penalties with a fixed numerical ratio of 9:1 (i.e., alpha = 0.9) and tuned for the Lambda value (i.e., the overall strength of regularization). For the random forest models, we used the *rfsrc()* function from *randomForestSRC* package for model training, where ntree = 1000 (i.e., Number of trees grown) and nsplit = 50 (i.e., the number of random splits to consider for each candidate splitting variable), and utilized 5-fold cross-validation to evaluate the predictive performance of the models. These procedures were repeated for each look back period and for each clinical setting. We report prediction metrics (i.e. area under the receiver operator characteristic curve (AUROC), as well as positive predictive value (PPV) and sensitivity with specificity set to 0.95 for model evaluation (Uno et al., 2007) using three prediction windows (t = 6, 12, and 18 months). The minimum length of prediction windows (6 months) was chosen to balance the need for sufficient outcome events with the desire to capture relatively short-term risk.

2.5. Sensitivity analysis: sampling all versus one random visit per patient

To empirically compare how our sampling scheme (i.e., randomly sampling one visit per patient) performs compared to the approach of including all visits, we performed sensitivity analyses by training the regularized Cox models with all patients visits available for each of the three cohorts. Models were fit to the all-visit data in the same manner as in the main analysis, except that to accommodate the larger number of visits, the down-sampling of the general outpatient cohort was done by first down-sampling the number of patients with no SRB events by a factor of 100 (instead of 20), which was then adjusted later via reweighting during model training. To maintain fairness to patients with less frequent visits, we also perform evaluation by restricting to sampling one visit per patient, i.e., during cross-validation, one visit was randomly sampled from patients who were not included in the inner splits.

3. Results

A total of 1,210,225, 13,098, and 7825 patients were included in the general outpatient, psychiatric-ED, and psychiatric inpatient cohorts, respectively. Table 1 summarizes the demographic composition of the three cohorts. The mean (SD) age for the three cohorts were 50.5 (18.0), 40.3 (17.4), and 43.4 (17.5). Overall, all cohorts comprised slightly more females compared to males (e.g. 60% females in the general cohort). Single marital status and public insurance were the majority status for both the psychiatric-ED (Single, 68%; public payer, 75%) and psychiatric inpatient cohorts (single, 60%; public payer, 68%), but not for the general cohort (single, 33%; public payer, 45%). Prevalence of suicide attempts was lowest in the general cohort (0.02%, 0.04% and 0.06% for 6 months, 12 months, and 18 months prediction windows) and highest in the psychiatric-ED cohort (1.6%, 2.3% and 3% for 6 months, 12 months, and 18 months prediction windows).

In general, both models showed good to excellent discrimination metrics (AUROC) for each cohort and across different prediction windows, though AUROC was generally higher for the Cox models compared to RF models. For example, for the general cohort, the Cox model had AUROC >=0.9 across all prediction windows Table 2, while the survival RF model had AUROC >= 0.84 across prediction windows. Henceforth we adopted the regularized Cox model as the main model (the metrics for the RF model are reported in the supplements). Performance metrics for the regularized Cox model are shown in Fig. 1 and for the RF model in supplementary Fig. 1. Table 2 provides exact model performance metrics for look back periods up to two years, stratified by each cohort for the regularized Cox model. Metrics for all look back periods for the regularized Cox model are provided in Supplementary Tables 1–3, and corresponding results for the RF model are provided in Supplementary Tables 4-6 Supple indows. At 95% specificity, the highest PPV (0.19) was observed in the psychiatric inpatient cohort using an 18 months prediction window. Although the outpatient setting has the lowest PPV, those classified as high risk are 12 times more likely to have SRB compared to the general population when predicted using

Table 1
Demographic composition of the three patient sets analyzed in this study.

Setting	General Outpatient		Psychiatric-ED		Psychiatric Inpatients	
	Mean	SD	Mean	SD	Mean	SD
Age	50.45	17.95	40.35	17.37	43.35	17.46
	N	%	N	%	N	%
Gender						
Female	721,976	59.66	6328	48.31	4234	54.11
Male	488,224	40.34	6769	51.68	3591	45.89
Unknown	25	0.00	1	0.01	0	0.00
Race						
Asian	55,703	4.60	403	3.08	229	2.93
Black	72,215	5.97	1360	10.38	672	8.59
White	953,723	78.81	10,103	77.13	6076	77.65
Other	86,611	7.16	825	6.30	565	7.22
Unknown	41,973	3.47	407	3.11	283	3.62
Ethnicity						
Hispanic	69,475	5.74	775	5.92	404	5.16
Non-Hispanic	1,140,750	94.26	12,323	94.08	7421	94.84
Marital status						
Single	398,131	32.90	8958	68.39	4671	59.69
Married	643,806	53.20	2090	15.96	1800	23.00
Partner	7043	0.58	86	0.66	42	0.54
Divorced	68,824	5.69	1122	8.57	756	9.66
Separated	11,356	0.94	287	2.19	145	1.85
Widowed	42,172	3.48	370	2.82	223	2.85
Other/Unknown	38,893	3.21	185	1.41	188	2.40
Veteran status						
Yes	56,720	4.69	543	4.15	274	3.50
No	986,126	81.48	11,542	88.12	6807	86.99
Unknown	167,379	13.83	1013	7.73	744	9.51
Public payer						
Yes	549,438	45.40	9861	75.29	5367	68.59
No	660,787	54.60	3237	24.71	2458	31.41
Total N	1,210,225		13,098		7825	

Table 2
Regularized Cox model performance metrics for the three clinical settings. (a) general outpatient; (b) psychiatric ED; (c) psychiatric inpatient. Prevalence is the proportion of visits followed by at least one SRB event within the specified prediction window and cohort. Thresholded metrics are reported at 95% specificity.

(a) Model metrics for general outpatient cohort						
Look back period (yrs)	Prediction window	AUROC	Sensitivity	PPV	Prevalence	Relative Risk
0.5	0.5	0.917	0.632	0.003	0.02%	12.612
1	0.5	0.923	0.648	0.003	0.02%	12.923
2	0.5	0.929	0.700	0.003	0.02%	13.954
0.5	1	0.914	0.624	0.005	0.04%	12.415
1	1	0.918	0.624	0.005	0.04%	12.414
2	1	0.925	0.669	0.005	0.04%	13.307
0.5	1.5	0.904	0.584	0.007	0.06%	11.597
1	1.5	0.907	0.577	0.007	0.06%	11.475
2	1.5	0.916	0.621	0.007	0.06%	12.325
(b) Model metrics for psychiatric ED cohort						
Look back period (yrs)	Prediction window	AUROC	Sensitivity	PPV	Prevalence	Relative Risk
0.5	0.5	0.742	0.278	0.084	1.62%	5.172
1	0.5	0.744	0.290	0.087	1.62%	5.381
2	0.5	0.739	0.302	0.091	1.62%	5.583
0.5	1	0.755	0.268	0.113	2.33%	4.858
1	1	0.761	0.282	0.119	2.33%	5.097
2	1	0.757	0.293	0.123	2.33%	5.271
0.5	1.5	0.769	0.309	0.160	3.00%	5.350
1	1.5	0.773	0.328	0.168	3.00%	5.630
2	1.5	0.769	0.324	0.167	3.00%	5.574
(c) Model metrics for psychiatric inpatient cohort						
Look back period (yrs)	Prediction window	AUROC	Sensitivity	PPV	Prevalence	Relative Risk
0.5	0.5	0.788	0.332	0.092	1.51%	6.110
1	0.5	0.782	0.332	0.093	1.51%	6.117
2	0.5	0.784	0.362	0.100	1.51%	6.612
0.5	1	0.796	0.386	0.148	2.20%	6.721
1	1	0.793	0.394	0.150	2.20%	6.834
2	1	0.795	0.394	0.150	2.20%	6.834
0.5	1.5	0.792	0.357	0.171	2.80%	6.102
1	1.5	0.790	0.362	0.173	2.80%	6.169
2	1.5	0.793	0.393	0.185	2.80%	6.602

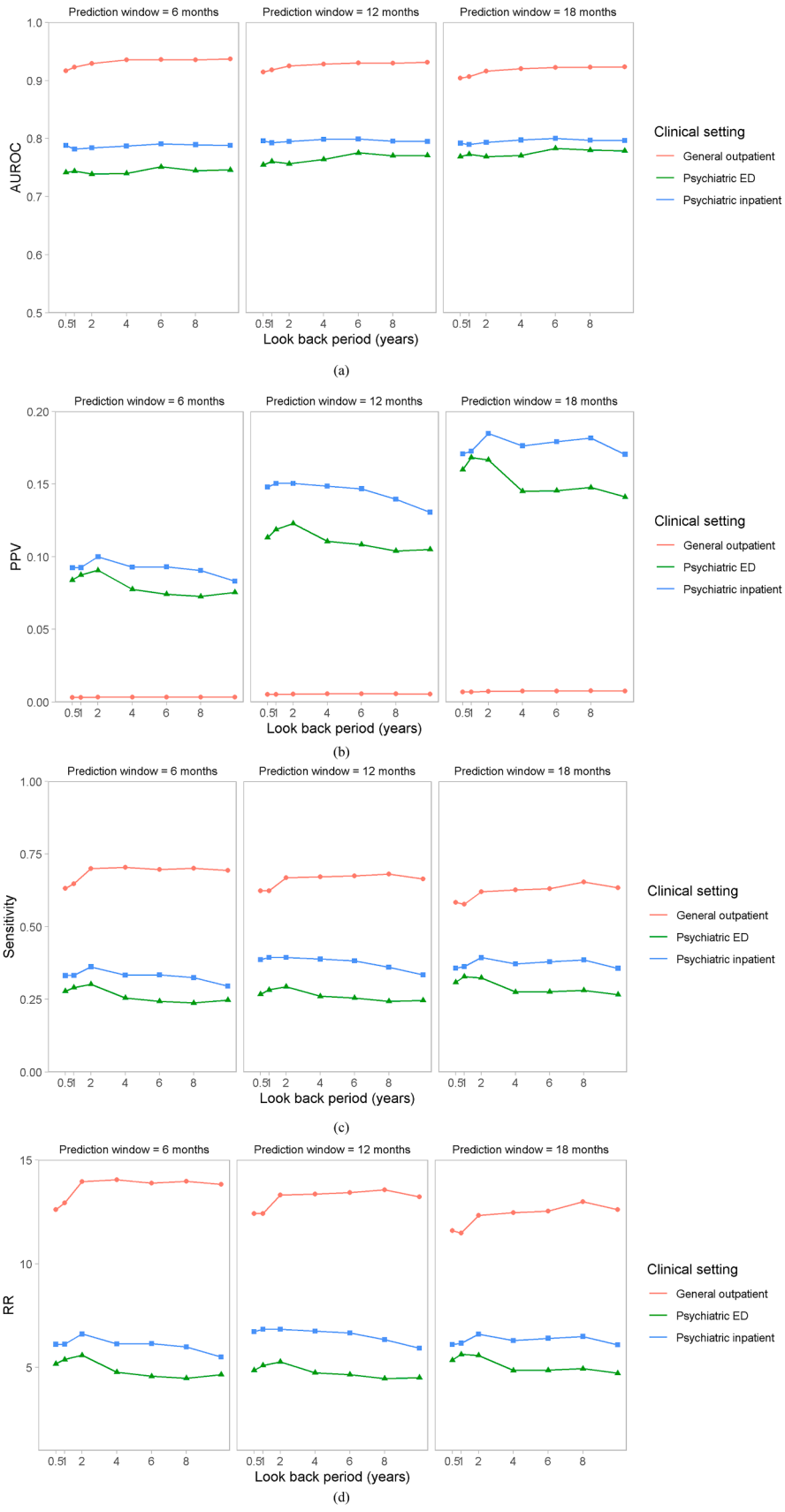

Fig. 1. Model metrics plots for regularized Cox regression. (a) AUROC; (b) PPV; (c) sensitivity; (d) relative risk (PPV/baseline prevalence of each cohort). Threshold-dependent metrics are reported at 95% specificity. For each model metric, each line plot (in rectangle) represents model metrics for specific prediction windows (6, 12, and 18 months, left to right). For each line plot, different lengths of look back period (x-axis) is plotted against the metric of interest (y-axis). Colors denote the clinical settings (red: general outpatient; green: psychiatric ED; blue: psychiatric inpatient). Thresholded metrics are plotted at 95% specificity. (a)AUROC; (b)PPV; (c) Sensitivity (d) Relative Risk (RR).

look back periods of 2 years or more look back window. As a general trend, model metrics were largely invariant to the length of look back period; using only recent (i.e., 6 months) of historical data provided comparable performance to using all data prior to the sampled visits. Fig. 2 plots the top 20 features by absolute value of beta coefficients for the regularized Cox model under each clinical setting with 1-year look back period (top features for other look back periods were similar within clinical setting). Given that the features are standardized, the beta coefficients are directly comparable.

Supplementary Tables 7–9 show model performance metrics for the regularized Cox models using all visits available. Despite utilizing more information, models trained using all visits performed on par with (e.g., AUROC 0.89–0.91 across all lookback periods for 6 months prediction window the general outpatient cohort) or slightly worse (e.g., AUROC 0.67–0.77 across all lookback periods for 6 months prediction window for the psychiatric ED and inpatient cohorts) than models trained with one visit per patient based on point estimates of the model metrics.

4. Discussion

We used large scale, structured EHR data to develop SRB prediction models using a landmark modeling approach that mirrors real-world scenarios in which a clinician would want to assess prospective risk of SRB. Both regularized Cox regression and RF survival models achieved high discrimination performance across all patient cohort/prediction window combinations. Our approach avoids temporal bias that arises from case-control sampling and provides an unbiased method for computing estimated risks across multiple clinical settings that is at the same time computationally efficient, which may be particularly beneficial in settings where computational resources are limited.

In recent years, the alarming frequency of suicide attempts and suicide-related deaths have underscored the urgent need to improve methods of risk stratification and prevention. The fact that most individuals who attempt or die by suicide are seen by healthcare providers in the preceding months provides a crucial opportunity for risk mitigation in clinical settings. In 2021, the US Surgeon General issued a "Call to Action to Implement the National Strategy for Suicide Prevention" (U.S. Department of Health and Human Services, 2021) that noted the value of EHRs for screening and assessment for suicide risk. There has recently been a proliferation of efforts to develop statistical and machine learning models that leverage the scale and breadth of EHR data for suicide risk prediction (Barak-Corren et al., 2017; Chen et al., 2020; Cho et al., 2020; Cohen et al., 2020; Karmakar et al., 2016; Levis et al., 2020; Nock et al., 2022; Nordin et al., 2021; Obeid et al., 2020; Simon et al., 2018; Su et al., 2020; Tsui et al., 2021; van Mens et al., 2020; Walsh et al., 2018; Xu et al., 2022; Zheng et al., 2020). Though such models have been shown to outperform risk assessment by clinicians (Nock et al., 2022), they have often had limitations including susceptibility to temporal bias, biases towards patients with more frequent visits, and, in some cases, intensive computational burden. In this context, the work described here includes several notable findings.

First, we demonstrate that models using a prospective landmark approach and structured EHR data showed good to excellent discrimination (AUROC = 0.74–0.93) across patient cohorts seen in general outpatient, psychiatric ED, and psychiatric inpatient care settings. In addition, by randomly sampling one visit per patient for each cohort, our approach to model training provides fairness to each patient regardless of their number of hospital visits, as well as greater computational efficiency compared to approaches that utilize all visits (Chen et al., 2020; Simon et al., 2018). However, performance metrics varied by clinical setting, likely due to differing base rates of suicide-related behavior. Models applied to the general outpatient cohort achieved the highest sensitivity (0.58–0.70 at 95% specificity) and discrimination, indexed by AUROC (0.90–0.93). At 95% specificity, patients classified as high-risk had a 12- to 13-fold increase risk of suicide attempt over a 6-month window. On the other hand, PPVs were lowest among the cohorts tested, ranging from 0.3 to 0.7%. For predictions from general outpatient settings, presumably reflecting the low base prevalence (e.g., 0.02% over 6 months). In contrast, for patients in the psychiatric ED cohort (where the base prevalence was 1.62% over 6 months), sensitivity was lowest (0.33–0.39) but PPVs were highest among the three care settings (8–17%). Results for the psychiatric inpatient cohort were

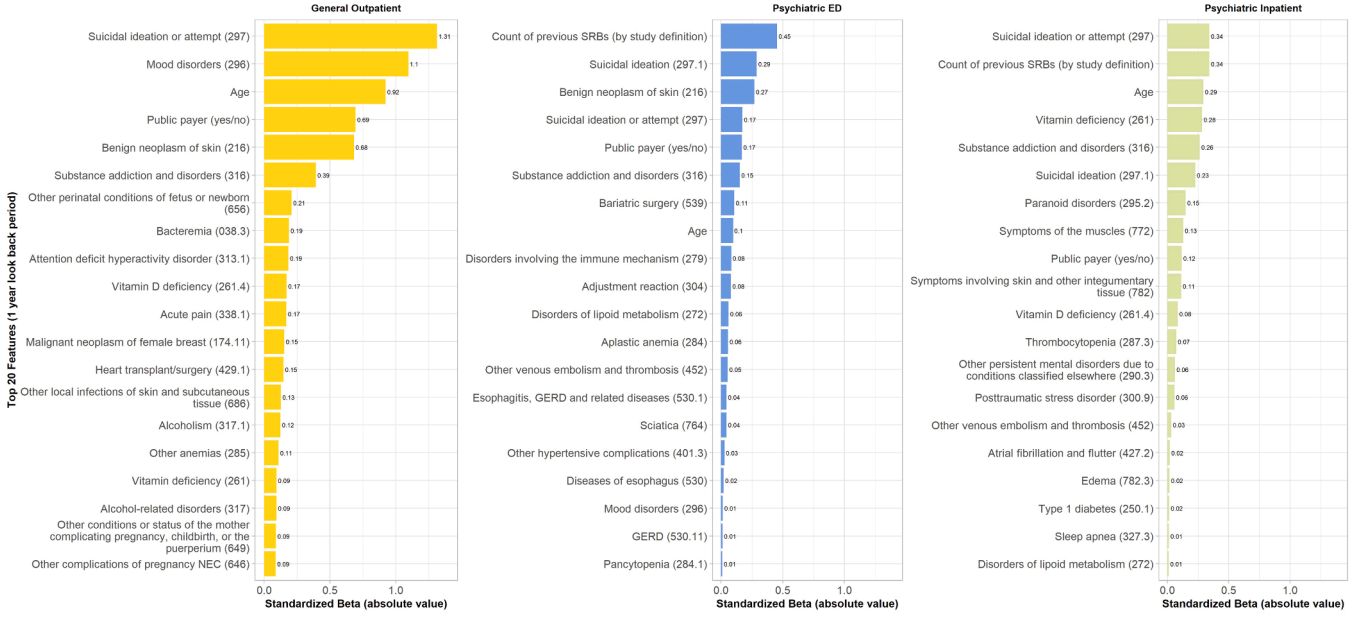

Fig. 2. Top 20 important features for regularized Cox regression, ranked by absolute value of beta coefficient of the regression model. Look back period = 1 year is provided here for illustrative purpose. Results for other look back periods are similar within each clinical setting. Number in the parenthesis represents the phecode for phecode-based features. Beta coefficients of the phecode features were calculated with per log number increase of normalized code counts. Note that all features are standardized to have a mean of 0 and standard deviation of 1, therefore the betas are directly comparable. The feature "Count of previous SRBs (by study definition)" refers to ICD codes that were included in the case definition; these codes were removed from other phecodes during feature engineering to avoid double counting. Public payer: whether the source of healthcare insurance was provided by a public source (e.g., Medicaid, Medicare).

intermediate between those seen for general outpatient and psychiatric ED cohorts. These differences in model performance are relevant for evaluating the net benefit of deploying such models for suicide risk screening in different contexts. That is, their utility will depend on whether the goal is to minimize false positives (which would favor use in the subset of patients seen in psychiatric EDs) or false negatives (in which case implementation in outpatient care settings could be justified). For the general outpatient cohort, false positives could be reduced by setting highly stringent thresholds for defining high-risk patients. Alternatively, improved performance might be achieved by incorporating additional data. However, in a set of analyzes during the earlier stages of this study (results not reported), we found that adding information such as medication, procedures, and features extracted via natural language processing did not substantially improve model performance. Recent work by our group (Nock et al., 2022) and others (Wilimitis et al., 2022) suggests that model performance can be enhanced by combining historical EHR data with in-person suicide risk survey data. Nevertheless, the model performance we observed in this study exceed thresholds we have previously shown to be cost-effective when paired with evidence-based prevention strategies (Ross et al., 2021).

Overall, our results compare favorably to those reported for prior EHR-based suicide risk models developed for prospective prediction (Chen et al., 2020; Cho et al., 2020; Karmakar et al., 2016; Simon et al., 2018; Walsh et al., 2021; Zheng et al., 2020). For example, Simon et al. (Simon et al., 2018) reported a PPV of approximately 2.5% over a 90-day window following a primary care visit with a mental health diagnosis and 5% following a mental health specialty visit, The AUROCs for these models (0.85) were comparable to those seen in the current study. Walsh and colleagues (Walsh et al., 2021) reported real-time predictions using a random forest model based on a 5-year look back period and incorporating EHR data and a zip code-based Area Deprivation Index. As in our study, discrimination metrics varied by clinical setting: health system-wide (AUROC = 0.84), ED (AUROC = 0.78), inpatient psychiatry (AUROC = 0.63). Model PPV for suicide attempt for the highest risk quantile was 0.4%.

Also of note, we find that model performance for a given prediction window is similar across look back periods from 2 years to as short as 6 months. This suggests that such models can be useful even when extensive historical data are not available for most patients. Similarly, for a given look back period, results were similar for prediction windows of 6 months to 18 months.

In the current study, model evaluation is reported based on one randomly sampled landmark visit per patient in the hold-out set during cross-validation. Another choice would be to treat all visits as landmark visits (i.e. visits from which predictions are made) in the model evaluation. However, our approach of selecting a set number of randomly sampled visits per patient provides equal sampling probability to each patient regardless of the frequency of their clinical visits. This avoids over-representation of patients with greater healthcare utilization. To empirically compare how our sampling scheme (i.e., randomly sampling one visit per patient) performs compared to the approach of including all visits, we performed sensitivity analyses by training the regularized Cox model using all visits available for each patient. Results show that models trained with all visits available did not perform better than their one-visit-per-patient counterparts. Theoretically, while models might benefit from the additional information in the extra visits, the repeated visits might not be as informative as expected, because the look-back windows (especially the longer ones) would already include much of the longitudinal patient history, even if only one visit was sampled per patient. Notably, in our study, using longer look-back windows did not improve performance meaningfully.

As mentioned, one additional advantage of the landmark approach is that by providing flexibility in sampling, the models can be more lightweight and efficient during training. In settings where more frequent updates are desired, training efficiency is valuable because training and tuning time for most machine learning models can grow as data size, features, and hyperparameters tuned increase (Efron et al., 2004; Keles et al., 2022; Sani et al., 2018), as do requirements for system memory. These demands can become problematic in circumstances involving big data and complex models, especially for resource-limited settings.

Our results should be interpreted in light of several limitations. First, because the models are based on EHR data, predictors and outcomes may not have been fully ascertained to the extent that patients receive care outside the health system. Second, the minimum prediction window evaluated here was 6 months. Further work will evaluate prediction windows (e.g. 30 days) that might be of clinical importance. Lastly, our models were trained and validated in a single healthcare system and performance in other systems may vary depending on variation in patient characteristics and documentation practices. Nevertheless, we note that the MGB system comprises more than 8 hospitals with heterogeneous catchment areas and clinical practices, supporting some degree of generalizability. In addition, the landmark framework used here is computationally efficient and can be readily used for local training and validation in other systems.

Integrating these models in day-to-day clinical practice imposes potential challenges that could be addressed in future work. These can include, and are not limited to: (1) balancing the cost of false positives/negatives; (2) issues around provider liability; (3) pairing predictions with evidence-based preventions; (4) education of clinicians and patients to ensure accurate interpretation of model predictions (e.g. to emphasize that risk scores are meant to inform, rather than replace, clinician judgment); and (5) appropriate design and implementation of clinical decision support systems (e.g. minimizing "alert fatigue").

In conclusion, we introduce a prospective landmark modeling approach using large-scale, structured EHR data that is accurate, reduces bias, and is computationally efficient for predicting suicide attempts in real-world healthcare settings. Given the urgency of improving risk stratification and prevention of suicide-related behavior, our approach offers a valuable opportunity for enhancing efforts to inform clinical decision making.

Ethics and consent statement

The authors assert that all procedures contributing to this work comply with the ethical standards of the relevant national and institutional committees on human experimentation and with the Helsinki Declaration of 1975, as revised in 2008. The study protocol was approved by the Institutional Review Board of Mass General Brigham (MGB) with a waiver of consent for the analysis of electronic health record data.

CRediT authorship contribution statement

Yi-han Sheu: Conceptualization, Methodology, Validation, Investigation, Data curation, Writing – original draft, Writing – review & editing, Visualization. **Jiehuan Sun:** Methodology, Software, Formal analysis, Investigation, Data curation, Writing – original draft, Writing – review & editing, Visualization. **Hyunjoon Lee:** Investigation, Data curation, Writing – review & editing. **Victor M. Castro:** Data curation, Writing – review & editing. **Yuval Barak-Corren:** Writing – review & editing. **Eugene Song:** Project administration, Writing – review & editing. **Emily M. Madsen:** Project administration, Writing – review & editing. **William J. Gordon:** Investigation, Writing – review & editing. **Isaac S. Kohane:** Conceptualization, Writing – review & editing. **Susanne E. Churchill:** Conceptualization, Project administration, Writing – review & editing. **Ben Y. Reis:** Methodology, Writing – review & editing. **Tianxi Cai:** Conceptualization, Methodology, Validation, Writing – review & editing, Supervision, Funding acquisition. **Jordan W. Smoller:** Conceptualization, Methodology, Resources, Validation, Writing – review & editing, Supervision, Funding acquisition.

Declaration of Competing Interest

Dr. Smoller is a member of the the Scientific Advisory Board of Sensorium Therapeutics (with equity), and has received grant support from Biogen, Inc. He is PI of a collaborative study of the genetics of depression and bipolar disorder sponsored by 23andMe for which 23andMe provides analysis time as in-kind support but no payments.

Acknowledgments

Funding: This work was supported in part by funding from Harvard Medical School (TC, JWS), the Tommy Fuss Fund, and NIMH R01 MH117599-05 (JWS).

Supplementary materials

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.psychres.2023.115175.

References

Ahmedani, B.K., Simon, G.E., Stewart, C., Beck, A., Waitzfelder, B.E., Rossom, R., Lynch, F., Owen-Smith, A., Hunkeler, E.M., Whiteside, U., Operskalski, B.H., Coffey, M.J., Solberg, L.I., 2014. Health care contacts in the year before suicide death. J. Gen. Intern. Med. 29, 870–877. https://doi.org/10.1007/s11606-014-2767-3.

Barak-Corren, Y., Castro, V.M., Javitt, S., Hoffnagle, A.G., Dai, Y., Perlis, R.H., Nock, M. K., Smoller, J.W., Reis, B.Y., 2017. Predicting suicidal behavior from longitudinal electronic health records. Am. J. Psychiatry 174, 154–162. https://doi.org/10.1176/appi.ajp.2016.16010077.

Barak-Corren, Y., Castro, V.M., Nock, M.K., Mandl, K.D., Madsen, E.M., Seiger, A., Adams, W.G., Applegate, R.J., Bernstam, E.V., Klann, J.G., McCarthy, E.P., Murphy, S.N., Natter, M., Ostasiewski, B., Patibandla, N., Rosenthal, G.E., Silva, G.S., Wei, K., Weber, G.M., Weiler, S.R., Reis, B.Y., Smoller, J.W., 2020. Validation of an electronic health record-based suicide risk prediction modeling approach across multiple health care systems. JAMA Netw. Open 3, e201262. https://doi.org/10.1001/jamanetworkopen.2020.1262.

Chen, Q., Zhang-James, Y., Barnett, E.J., Lichtenstein, P., Jokinen, J., D'Onofrio, B.M., Faraone, S.V., Larsson, H., Fazel, S., 2020. Predicting suicide attempt or suicide death following a visit to psychiatric specialty care: a machine learning study using Swedish national registry data. PLoS Med. 17, e1003416 https://doi.org/10.1371/journal.pmed.1003416.

Cho, S.E., Geem, Z.W., Na, K.S., 2020. Prediction of suicide among 372,813 individuals under medical check-up. J. Psychiatr. Res. 131, 9–14. https://doi.org/10.1016/j.jpsychires.2020.08.035.

Cohen, J., Wright-Berryman, J., Rohlfs, L., Wright, D., Campbell, M., Gingrich, D., Santel, D., Pestian, J., 2020. A feasibility study using a machine learning suicide risk prediction model based on open-ended interview language in adolescent therapy sessions. Int. J. Environ. Res. Public Health 17. https://doi.org/10.3390/ijerph17218187.

Denny, J.C., Bastarache, L., Ritchie, M.D., Carroll, R.J., Zink, R., Mosley, J.D., Field, J.R., Pulley, J.M., Ramirez, A.H., Bowton, E., Basford, M.A., Carrell, D.S., Peissig, P.L., Kho, A.N., Pacheco, J.A., Rasmussen, L.V., Crosslin, D.R., Crane, P.K., Pathak, J., Bielinski, S.J., Pendergrass, S.A., Xu, H., Hindorff, L.A., Li, R., Manolio, T.A., Chute, C.G., Chisholm, R.L., Larson, E.B., Jarvik, G.P., Brilliant, M.H., McCarty, C.A., Kullo, I.J., Haines, J.L., Crawford, D.C., Masys, D.R., Roden, D.M., 2013. Systematic comparison of phenome-wide association study of electronic medical record data and genome-wide association study data. Nat. Biotechnol. 31, 1102–1110. https://doi.org/10.1038/nbt.2749.

Efron, B., Hastie, T., Johnstone, I., Tibshirani, R., 2004. Least angle regression. Ann. Stat. 32, 407–499. https://doi.org/10.1214/009053604000000067.

Fisher, L.B., Curtiss, J.E., Klyce, D.W., Perrin, P.B., Juengst, S.B., Gary, K.W., Niemeier, J. P., Hammond, F.M., Bergquist, T.F., Wagner, A.K., Rabinowitz, A.R., Giacino, J.T., Zafonte, R.D., 2022. Using machine learning to examine suicidal ideation after TBI: a TBI model systems national database study. Am. J. Phys. Med. Rehabil. https://doi.org/10.1097/PHM.0000000000002054.

Ishwaran, H., Kogalur, U.B., Blackstone, E.H., Lauer, M.S., 2008. Random survival forests. Ann. Appl. Stat. 2, 841–860. https://doi.org/10.1214/08-AOAS169.

Karmakar, C., Luo, W., Tran, T., Berk, M., Venkatesh, S., 2016. Predicting risk of suicide attempt using history of physical illnesses from electronic medical records. JMIR Ment. Health 3, e19. https://doi.org/10.2196/mental.5475.

Keles F.D., Wijewardena P.M., Hegde C., 2022. On the Computational Complexity of Self-Attention. arXiv. doi:10.48550/arxiv.2209.04881.

Levis, M., Leonard Westgate, C., Gui, J., Watts, B.V., Shiner, B., 2020. Natural language processing of clinical mental health notes may add predictive value to existing suicide risk models. Psychol. Med. 1–10. https://doi.org/10.1017/S0033291720000173.

Levis, M., Levy, J., Dufort, V., Gobbel, G.T., Watts, B.V., Shiner, B., 2022. Leveraging unstructured electronic medical record notes to derive population-specific suicide risk models. Psychiatry Res. 315, 114703 https://doi.org/10.1016/j.psychres.2022.114703.

Machado, C.D.S., Ballester, P.L., Cao, B., Mwangi, B., Caldieraro, M.A., Kapczinski, F., Passos, I.C., 2022. Prediction of suicide attempts in a prospective cohort study with a nationally representative sample of the US population. Psychol. Med. 52, 2985–2996. https://doi.org/10.1017/S0033291720004997.

Malone, T.L., Zhao, Z., Liu, T.Y., Song, P.X.K., Sen, S., Scott, L.J., 2021. Prediction of suicidal ideation risk in a prospective cohort study of medical interns. PLoS One 16, e0260620. https://doi.org/10.1371/journal.pone.0260620.

Nalichowski, R., Keogh, D., Chueh, H.C., Murphy, S.N., 2006. Calculating the benefits of a research patient data repository. AMIA Annu. Symp. Proc. 1044.

National Center for Health Statistics, 2018. Suicide Mortality in the United States, 1999–2017. NCHS Data Brief 330.

Nock, M.K., Millner, A.J., Ross, E.L., Kennedy, C.J., Al-Suwaidi, M., Barak-Corren, Y., Castro, V.M., Castro-Ramirez, F., Lauricella, T., Murman, N., Petukhova, M., Bird, S. A., Reis, B., Smoller, J.W., Kessler, R.C., 2022. Prediction of suicide attempts using clinician assessment, patient self-report, and electronic health records. JAMA Netw. Open 5, e2144373. https://doi.org/10.1001/jamanetworkopen.2021.44373.

Nordin, N., Zainol, Z., Mohd Noor, M.H., Lai Fong, C., 2021. A comparative study of machine learning techniques for suicide attempts predictive model. Health Inform. J. 27, 1460458221989395 https://doi.org/10.1177/1460458221989395.

Obeid, J.S., Dahne, J., Christensen, S., Howard, S., Crawford, T., Frey, L.J., Stecker, T., Bunnell, B.E., 2020. Identifying and predicting intentional self-harm in electronic health record clinical notes: deep learning approach. JMIR Med. Inform. 8, e17784. https://doi.org/10.2196/17784.

Parast, L., Cai, T., 2013. Landmark risk prediction of residual life for breast cancer survival. Stat. Med. 32, 3459–3471. https://doi.org/10.1002/sim.5776.

Parast, L., Cheng, S.C., Cai, T., 2012. Landmark prediction of long term survival incorporating short term event time information. J. Am. Stat. Assoc. 107, 1492–1501. https://doi.org/10.1080/01621459.2012.721281.

Rossom, R.C., Yarborough, B.J., Boggs, J.M., Coleman, K.J., Ahmedani, B.K., Lynch, F.L., Daida, Y., Simon, G.E., 2021. Prediction of suicidal behavior using self-reported suicidal ideation among patients with bipolar disorder. J. Affect. Disord. 295, 410–415. https://doi.org/10.1016/j.jad.2021.08.060.

Ross, E.L., Zuromski, K.L., Reis, B.Y., Nock, M.K., Kessler, R.C., Smoller, J.W., 2021. Accuracy requirements for cost-effective suicide risk prediction among primary care patients in the US. JAMA Psychiatry 78, 642–650. https://doi.org/10.1001/jamapsychiatry.2021.0089.

Sani, H.M., Lei, C., Neagu, D., Bramer, M., Petridis, M., 2018. Computational complexity analysis of decision tree algorithms. Eds.. In: Proceedings of the Artificial Intelligence XXXV: 38th SGAI International Conference on Artificial Intelligence. Springer International Publishing, Cham, pp. 191–197. https://doi.org/10.1007/978-3-030-04191-5_17. AI 2018, Cambridge, UK, December 11–13, 2018, Proceedings, Lecture Notes in Computer Science.

Simon, G.E., Johnson, E., Lawrence, J.M., Rossom, R.C., Ahmedani, B., Lynch, F.L., Beck, A., Waitzfelder, B., Ziebell, R., Penfold, R.B., Shortreed, S.M., 2018. Predicting suicide attempts and suicide deaths following outpatient visits using electronic health records. Am. J. Psychiatry 175, 951–960. https://doi.org/10.1176/appi.ajp.2018.17101167.

Simon, G.E., Yarborough, B.J., Rossom, R.C., Lawrence, J.M., Lynch, F.L., Waitzfelder, B. E., Ahmedani, B.K., Shortreed, S.M., 2019. Self-reported suicidal ideation as a predictor of suicidal behavior among outpatients with diagnoses of psychotic disorders. Psychiatr. Serv. 70, 176–183. https://doi.org/10.1176/appi.ps.201800381.

Simon, N., Friedman, J., Hastie, T., Tibshirani, R., 2011. Regularization paths for Cox's proportional hazards model via coordinate descent. J. Stat. Softw. 39, 1–13. https://doi.org/10.18637/jss.v039.i05.

Su, C., Aseltine, R., Doshi, R., Chen, K., Rogers, S.C., Wang, F., 2020. Machine learning for suicide risk prediction in children and adolescents with electronic health records. Transl. Psychiatry 10, 413. https://doi.org/10.1038/s41398-020-01100-0.

Tsui, F.R., Shi, L., Ruiz, V., Ryan, N.D., Biernesser, C., Iyengar, S., Walsh, C.G., Brent, D. A., 2021. Natural language processing and machine learning of electronic health records for prediction of first-time suicide attempts. JAMIA Open 4. https://doi.org/10.1093/jamiaopen/ooab011 ooab011.

U.S. Department of Health and Human Services, 2021. The surgeon general releases call to action to implement [WWW Document]. URL https://www.hhs.gov/about/news/2021/01/19/the-surgeon-general-releases-call-to-action-to-implement-the-national-strategy-for-suicide-prevention.html (accessed 8.19.22).

Uno, H., Cai, T., Tian, L., Wei, L.J., 2007. Evaluating prediction rules fort -year survivors with censored regression models. J. Am. Stat. Assoc. 102, 527–537. https://doi.org/10.1198/016214507000000149.

Van Houwelingen, H.C., 2007. Dynamic prediction by landmarking in event history analysis. Scand. J. Stat. 34, 70–85. https://doi.org/10.1111/j.1467-9469.2006.00529.x.

van Mens, K., Elzinga, E., Nielen, M., Lokkerbol, J., Poortvliet, R., Donker, G., Heins, M., Korevaar, J., Dückers, M., Aussems, C., Helbich, M., Tiemens, B., Gilissen, R., Beekman, A., de Beurs, D., 2020. Applying machine learning on health record data from general practitioners to predict suicidality. Internet Interv. 21, 100337 https://doi.org/10.1016/j.invent.2020.100337.

Walsh, C.G., Johnson, K.B., Ripperger, M., Sperry, S., Harris, J., Clark, N., Fielstein, E., Novak, L., Robinson, K., Stead, W.W., 2021. Prospective validation of an electronic health record-based, real-time suicide risk model. JAMA Netw. Open 4, e211428. https://doi.org/10.1001/jamanetworkopen.2021.1428.

Walsh, C.G., Ribeiro, J.D., Franklin, J.C., 2017. Predicting risk of suicide attempts over time through machine learning. Clin. Psychol. Sci. 5, 457–469. https://doi.org/10.1177/2167702617691560.

Walsh, C.G., Ribeiro, J.D., Franklin, J.C., 2018. Predicting suicide attempts in adolescents with longitudinal clinical data and machine learning. J. Child Psychol. Psychiatry 59, 1261–1270. https://doi.org/10.1111/jcpp.12916.

Wei, Y.X., Liu, B.P., Zhang, J., Wang, X.T., Chu, J., Jia, C.X., 2021. Prediction of recurrent suicidal behavior among suicide attempters with Cox regression and machine learning: a 10-year prospective cohort study. J. Psychiatr. Res. 144, 217–224. https://doi.org/10.1016/j.jpsychires.2021.10.023.

Wilimitis, D., Turer, R.W., Ripperger, M., McCoy, A.B., Sperry, S.H., Fielstein, E.M., Kurz, T., Walsh, C.G., 2022. Integration of face-to-face screening with real-time machine learning to predict risk of suicide among adults. JAMA Netw. Open 5, e2212095. https://doi.org/10.1001/jamanetworkopen.2022.12095.

World Health Organization, 1993. ICD-10 : international statistical classification of diseases and related health problems : tenth revision.

Wu, Y., 2012. Elastic net for Cox's proportional hazard model with a solution path algorithm. Stat. Sin. 22, 27–294. https://doi.org/10.5705/ss.2010.107.

Xu, W., Su, C., Li, Y., Rogers, S., Wang, F., Chen, K., Aseltine, R., 2022. Improving suicide risk prediction via targeted data fusion: proof of concept using medical claims data. J. Am. Med. Inform. Assoc. 29, 500–511. https://doi.org/10.1093/jamia/ocab209.

Yuan, W., Beaulieu-Jones, B.K., Yu, K.H., Lipnick, S.L., Palmer, N., Loscalzo, J., Cai, T., Kohane, I.S., 2021. Temporal bias in case-control design: preventing reliable predictions of the future. Nat. Commun. 12, 1107. https://doi.org/10.1038/s41467-021-21390-2.

Zheng, L., Wang, O., Hao, S., Ye, C., Liu, M., Xia, M., Sabo, A.N., Markovic, L., Stearns, F., Kanov, L., Sylvester, K.G., Widen, E., McElhinney, D.B., Zhang, W., Liao, J., Ling, X. B., 2020. Development of an early-warning system for high-risk patients for suicide attempt using deep learning and electronic health records. Transl. Psychiatry 10, 72. https://doi.org/10.1038/s41398-020-0684-2.

Zheng, W., Jiang, M.L., He, H.B., Li, R.P., Li, Q.L., Zhang, C.P., Zhou, S.M., Yan, S., Ning, Y.P., Huang, X., 2020. Serum BDNF levels are not associated with the antidepressant effects of nonconvulsive electrotherapy. Neuropsychiatr. Dis. Treat. 16, 1555–1560. https://doi.org/10.2147/NDT.S256278.