Importance of variables from different time frames for predicting self-harm using health system data

Charles J. Wolock, PhD[1]; Brian D. Williamson, PhD[2,3]; Susan M. Shortreed, PhD[2,3]; Gregory E. Simon, MD, MPH[2,4]; Karen J. Coleman, PhD[4,5]; Rodney Yeargans[5]; Brian K. Ahmedani, MSW, PhD[6]; Yihe Daida, PhD[7]; Frances L. Lynch, PhD, MSPH[8]; Rebecca C. Rossom, MD, MS[9]; Rebecca A. Ziebell, BS[2]; Maricela Cruz, PhD[2,3]; Robert D. Wellman, MS[2]; R. Yates Coley, PhD[2,3]

[1]Department of Biostatistics, Epidemiology and Informatics, University of Pennsylvania
[2]Kaiser Permanente Washington Health Research Institute
[3]Department of Biostatistics, University of Washington
[4]Department of Health Systems Science, Bernard J. Tyson Kaiser Permanente School of Medicine
[5]Department of Research and Evaluation, Kaiser Permanente Southern California
[6]Center for Health Policy and Health Services Research, Henry Ford Health
[7]Center for Integrated Health Care Research, Kaiser Permanente Hawaii
[8]Center for Health Research, Kaiser Permanente Northwest
[9]HealthPartners Institute

Corresponding author:
Charles J. Wolock, PhD
Department of Biostatistics, Epidemiology and Informatics
University of Pennsylvania
423 Guardian Drive, Philadelphia, PA 19104, USA
cwolock@upenn.edu

Keywords:
clinical prediction models
feature importance
insurance claims data
predictive analytics
suicide

ABSTRACT

Objective: Self-harm risk prediction models developed using health system data (electronic health records and insurance claims information) often use patient information from up to several years prior to the index visit when the prediction is made. Measurements from some time periods may not be available for all patients. We study the predictive potential of variables corresponding to different time horizons prior to the index visit.

Materials and Methods: We use variable importance to quantify the potential of recent (up to three months before the index visit) and distant (more than one year before the index visit) patient mental health information for predicting self-harm risk using data from seven health systems. We quantify importance as the decrease in predictiveness when the variable set of interest is excluded from the prediction task. We define predictiveness using discriminative metrics: area under the receiver operating characteristic curve (AUC), sensitivity, and positive predictive value.

Results: Mental health predictors corresponding to the three months prior to the index visit show strong signal of importance; in one setting, excluding these variables decreased AUC from 0.85 to 0.77. Predictors corresponding to more distant information were less important.

Discussion: Predictors from the months immediately preceding the index visit are highly important. Implementation of self-harm prediction models may be challenging in settings where recent data are not completely available (e.g., due to lags in insurance claims processing) at the time a prediction is made.

Conclusion: Clinically derived variables from different time frames exhibit varying levels of importance for predicting self-harm.

BACKGROUND

Preventing fatal and non-fatal self-harm is a public health priority: In the United States alone, in 2021 over 48,000 people died by suicide and an estimated 1.7 million adults attempted suicide [1, 2]. Health care settings provide an opportunity to prevent self-harm behavior if those at higher risk can be accurately identified. Health system data, including electronic health records (EHR) and health insurance claims data, contain detailed clinical history relevant to mental health risk factors and other predictors of self-harm. Additionally, identification of self-harm risk using health system data enables implementation of risk prediction models within EHR platforms for clinical use. Several models have been developed using health system data to predict the risk of fatal and non-fatal self-harm [3–14]. Many of these prediction models achieve an area under the receiver operating characteristic curve (AUC) of 0.8 or above, implying good prediction performance as measured by discrimination between those who do and those who do not attempt or die by suicide.

Self-harm risk varies over time; suicidal ideation, depressive symptoms, and other factors are not static. Many clinical prediction models for self-harm risk use information available prior to a medical visit to assess a patient's risk. We refer to the visit at which the prediction is made as the *index visit*. For predictors that can vary over time, it is common for these models to use information from up to five years prior to an index visit [7, 14–17]. In most cases, the predictors are divided into several overlapping time intervals: for example, predictors corresponding to information from the 90 days, one year, and five years prior to the visit [14, 15, 17], or 30 days, 90 days, and one year prior to the visit [6].

Information measured close in time to a given visit, including recent diagnoses and dispensed prescriptions, is likely correlated with an individual's current risk. However, it can be difficult to incorporate recent information into prediction models in real time. For example, there are often time lags in processing pharmacy or insurance claims data, especially from external providers, which means these data might not be available for risk prediction at the time of the index visit. Since predictors are often defined as the presence of a particular event, if the relevant data has not been processed, the predictor value is coded as zero, representing the absence of an event. For example, in a health system without inpatient facilities, a patient with an inpatient hospitalization in the past three months may have no record of this at an index visit because the claim has not yet

been processed. This lag may be a concern when deploying a risk prediction model. An additional concern is that some people have shorter clinical history available for prediction because they are new to the health system; health care information prior to enrollment in the current health system may be invisible to a prediction model. Shorter duration of clinical history could reflect less access to continuous insurance coverage and could be related to social determinants of health. Thus, it is of interest to determine the predictiveness of variables from particular time frames.

The statistical framework of *variable importance* can be used to investigate the predictiveness of a variable or group of variables. Variable importance can be broadly classified as either specific or agnostic to the algorithm used to construct the prediction model. Algorithm-specific variable importance measures (VIMs) quantify how the particular fitted algorithm uses variables to make predictions. Examples include the Gini criteria VIM returned by random forests algorithms [18], coefficients in penalized regression models [19], and changes in the prediction output by a model when certain variables are treated as missing [see, e.g., 20]. Algorithm-agnostic variable importance, in contrast, is the change in population prediction performance when certain variables are excluded from the model [see, e.g., 21–23]. Because algorithm-agnostic importance is not tied to a particular modeling strategy, its interpretation does not depend on the prediction technique used. Furthermore, by treating variable importance as a population quantity, the algorithm-agnostic approach allows for statistical inference. Both types of variable importance can provide complementary information [24]. However, the chosen VIM should reflect the scientific question at hand.

OBJECTIVE

Our goal in this work is to understand the implications of using different subsets of temporally defined variables to develop prediction models for fatal and non-fatal self-harm using health system data. In particular, we aim to quantify improvement in self-harm prediction attributed to the inclusion of clinically derived variables from more recent (0–3 months) and distant (13–60 months) time periods preceding the index visit. We judged the algorithm-agnostic variable importance approach to be most appropriate for this task, because our goal is to understand the implications of using different subsets of variables to develop prediction models, rather than to understand how any given prediction model makes use of the variables it is provided.

MATERIALS AND METHODS

Setting and study sample

Data for this study were collected from seven integrated health systems (HealthPartners, Henry Ford Health, and the Colorado, Hawaii, Northwest, Southern California, and Washington regions of Kaiser Permanente) that provide insurance coverage and comprehensive medical care to defined patient populations. Health system data, including EHR and insurance claims data, were extracted via each site's research data warehouse [25]. Responsible Institutional Review Boards for each health system granted waivers of consent to use de-identified records data for this research.

The study sample included two categories of visits made by members aged 11 or older: mental health specialty visits (mental health setting) and general medical visits in which a mental health diagnosis was recorded (general medical setting). There were no eligibility requirements related to prior health insurance plan enrollment or health care utilization. All eligible visits from January 1, 2009 to September 30, 2017 were included in the study sample with the exception of visits from Henry Ford Health, which only contributed data following the implementation of a new electronic records system on January 1, 2012. Members may have had multiple eligible visits in either the mental health or general medical setting during this time period and, as such, multiple visits per member were included in the analytic sample.

Outcomes and follow-up

We separately considered prediction of two binary outcomes: any self-harm (including fatal and non-fatal) and suicide death (i.e., fatal self-harm only) within 90 days of an eligible visit. As in prior work, non-fatal self-harm events were ascertained from health system data, either the EHR or claims data, by identifying all ICD-9/10 injury or poisoning diagnoses accompanied by a cause of injury code indicating intentional self-harm or undetermined intent [8]. Suicide deaths were ascertained from state mortality records and were identified by a cause-of-death code for self-inflicted injury or injury or poisoning with undetermined intent [26, 27]. Diagnosis code lists are available online at `https://github.com/MHResearchNetwork/more-srpm/blob/main/SRS3_DX_CODES_20181204.sas`.

Prediction models for any self-harm excluded patients who were not enrolled in the health

system's insurance plan on the index date or for the following 90 days to enable complete outcome capture from insurance claims data. Insurance plan enrollment was not required for inclusion in suicide death prediction models, as mortality records data were available on all patients regardless of their current enrollment status. The study sample for suicide death prediction excluded visits that occurred after availability of cause-of-death data at each site (Supplementary Table S1).

Self-harm risk predictors

Predictors of self-harm were extracted from EHR and insurance billing information. 'Base' predictors included in all prediction models, but not assessed for variable importance, included age, sex, race, ethnicity, insurance type, and census-derived sociodemographic variables.

All other predictors were related to mental health diagnoses or mental health care utilization. Mental health–specific predictors covering the 0–3 months, 4–12 months, and 13–60 months prior to the visit included binary indicators of the following in each time period: mental health and substance use diagnoses (e.g., depression, anxiety, alcohol use disorder); dispensed psychiatric medications (e.g., antidepressants, benzodiazepines); prior outpatient, inpatient, and emergency department encounters with mental health diagnoses; prior suicide attempt and self-harm diagnoses; and responses to the Patient Health Questionnaire 9th item [3, 15, 28, 29], which asks about suicidal ideation. A complete list of predictors is given in Supplementary Table S2.

Statistical analysis

We assessed variable importance by computing the difference in predictiveness (see predictiveness measures below) between pairs of fitted prediction models. The models considered are depicted in Figure 1. For a given predictor group of interest, the difference in predictiveness between a *larger model* (which includes the predictor group of interest and others) and *reduced model* (which excludes the predictor group of interest) quantifies how much predictiveness is lost by excluding predictors measured in a particular time period. This decrease in predictiveness is the variable importance of the predictor group relative to the full set of predictors in the larger model. Each model included the base predictors and some combination of month 0–3 predictors, month 4–12 predictors, and month 13–60 predictors. The models compared were:

- a model using months 0–12 was compared to a model using months 4–12 to assess the importance of month 0–3 predictors relative to month 0–12 predictors;
- a model using months 0–60 was compared to a model using months 4–60 to assess the importance of month 0–3 predictors relative to month 0–60 predictors;
- a model using months 4–60 was compared to a model using months 4–12 to assess the importance of month 13–60 predictors relative to month 4–60 predictors; and
- a model using months 0–60 was compared to a model using months 0–12 to assess the importance of month 13–60 predictors relative to month 0–60 predictors.

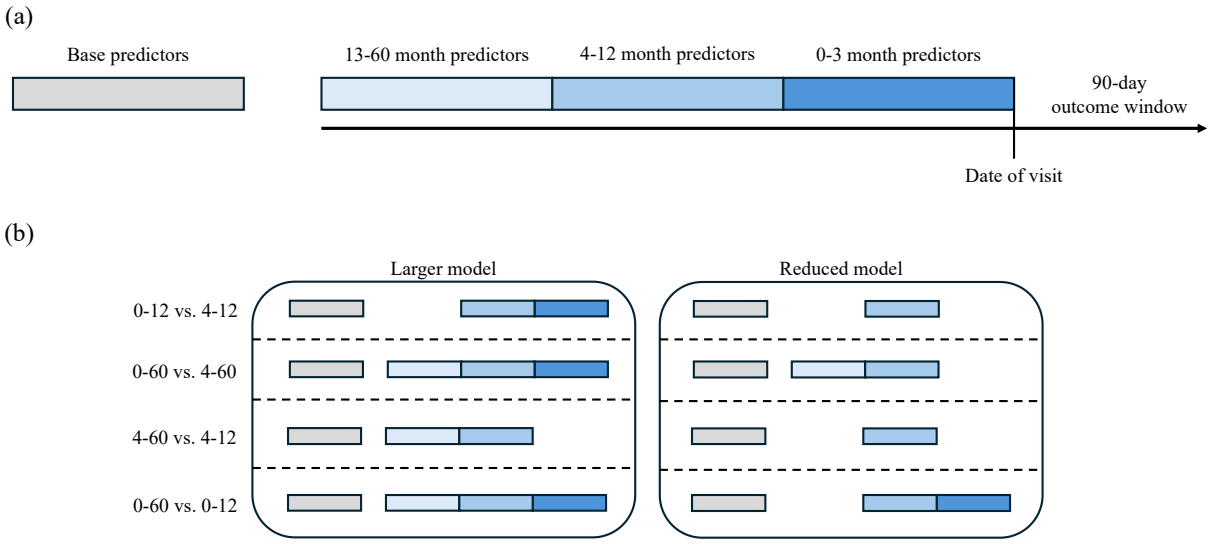

Figure 1: Schematic of temporal predictor groups in the variable importance analysis. (a) Predictors were categorized into four groups: base predictors, including demographics and comorbidities, that were included in all prediction models (gray), and mental health–specific predictors covering the 0–3 months (dark blue), 4–12 months (medium blue), and 13–60 months (light blue) prior to the prediction instance (vertical black line). The outcome window spanned 90 days from the prediction instance (date of the visit). Note that the timeline is not drawn to scale. (b) We made four comparisons to assess variable importance. In each case, the larger model used base predictors plus some subset of temporal predictors. The reduced model was constructed by removing a temporal predictor group.

Variable importance was quantified using predictiveness measures corresponding to AUC [30], sensitivity, and positive predictive value (PPV). Sensitivity and PPV were calculated using cut-points based on the 90^{th}, 95^{th}, and 99^{th} percentiles of predicted risk. These quantities measure discriminative performance. This choice of predictiveness measures accords with the practical usage of the models to identify a subset of patients at higher risk; the methods described here could be applied using other measures.

We assessed variable importance in four setting-outcome pairs defined by visit type (mental health specialty vs. general medical) and outcome (any self-harm event vs. suicide death). For each, a sample of patient-visits was constructed according to visit type and outcome-specific inclusion criteria (described above). For each of the four outcome-setting pairs, we performed the procedure described below. Our approach leverages a cross-fitting procedure with validation of model performance in independent testing sets. This combination of techniques reduces bias due to potential overfitting of prediction models and enables robust inference on variable importance [23]. The procedure was as follows:

1. The full collection of patients contributing visits to each sample was randomly subdivided on the person level into five folds, which we refer to as *cross-fitting* folds. Cross-fitting, which entails training the prediction model and evaluating variable importance on separate subsets of data, has been shown to improve performance in variable importance analyses [23, 31].

2. Three cross-fitting folds were designated as training data; the remaining two were designated as test data.

3. Visits corresponding to patients in the training data were used to construct penalized logistic regression models via the lasso [32]. The lasso is a regression method that combines shrinkage of coefficients towards zero and exclusion of variables with estimated null coefficients from the prediction model. The lasso penalization parameter was selected via 10-fold cross-validation [33] within the training data (the set of three cross-fitting folds from step 2) using AUC loss, with cross-validation folds defined on the person level (rather than the visit level) to ensure independence between folds.

4. The fitted lasso model was used to generate cross-fit predicted probabilities for the visits corresponding to patients in the test data (the set of two cross-fitting folds from step 2).

5. Sample-split predictiveness estimates of the two models being compared were computed separately on the two independent folds comprising the test sample. This permits valid inference even under zero importance, as evaluating both models in the same sample risks type I error inflation [23, 34]. VIM estimates, given by the difference in estimated predictiveness of the

two models, were truncated at zero. (The VIM parameter takes non-negative values by construction; increasing the number of available predictors cannot reduce model performance on a population level.) Variance estimates for predictiveness were computed separately on the two test folds using the nonparametric bootstrap [35] with 500 bootstrap replicates, resampled at the patient level. The variance of the VIM estimator was computed as the sum of variance estimates constructed from the two independent test folds [23].

6. To increase robustness against the random splitting of the data into folds, Steps 2–5 were repeated ten times, with different combinations of cross-fitting folds designed as training and test data. VIM estimates and corresponding variance estimates were averaged over all ten test/train combinations to give the final results. The final variance estimates were used to construct 95% confidence intervals based on a normal approximation.

RESULTS

A total of 15,986,946 mental health visits made by 1,590,002 patients and 11,104,580 general medical visits made by 2,732,786 patients were included in our analysis. The rates of self-harm and suicide death in the 90 days following a visit were 0.64% and 0.023%, respectively, in the mental health sample and 0.33% and 0.016% in the general medical sample. Overall characteristics of the study sample are summarized in Table 1, and a subset of the temporal predictors are summarized in Table 2.

Performance of risk prediction models using predictors from all time periods is reported in Table 3. The AUC estimates range from 0.807 to 0.850, with superior performance observed for predicting any self-harm versus suicide death. We observe a similar pattern for sensitivity. For example, sensitivity using the 95th risk score percentile cut-point was 45.7% for predicting any self-harm following a mental health specialty visit and 47.2% for predicting any self-harm following a general medical visit; for predicting suicide death, these values were 35.1% and 38.4%, respectively. PPV for predicting suicide death was low, in keeping with the low prevalence of fatal self-harm. Overall, these estimates are similar to those observed in previous studies of self-harm risk prediction in this setting [8], suggesting that the fitted lasso models perform as expected.

Figures 2–4 show the variable importance results for predictiveness measures corresponding

	Mental health $n = 15,986,946$	General medical $n = 11,104,580$
	n (%)	n (%)
Female	10,173,202 (63.6%)	6,964,601 (62.7%)
Age, years		
11–17	1,762,956 (11.0%)	692,306 (6.2%)
18–29	2,700,008 (16.9%)	1,418,491 (12.8%)
30–44	4,068,753 (25.5%)	2,191,953 (19.7%)
45–64	5,601,992 (35.0)	3,832,204 (34.5%)
65 and older	1,853,237 (11.6%)	2,969,626 (26.7%)
Race, self-reported		
American Indian/Alaskan Native	152,863 (0.96)	125,382 (1.1%)
Asian	785,358 (4.9%)	522,750 (4.7%)
Black/African American	1,393,712 (8.7%)	868,921 (7.8%)
Native Hawaiian/Pacific Islander	169,010 (1.1%)	101,295 (0.91%)
White	10,858,962 (67.9%)	7,808,508 (70.3%)
Multiple or other races indicated	85,075 (0.53%)	95,466 (0.86%)
Hispanic/Latino ethnicity	3,876,798 (24.2%)	2,384,030 (21.5%)
No race or ethnicity recorded	615,203 (3.8%)	384,300 (3.5%)
Insurance[a]		
Commercial group	11,669,276 (73.0%)	6,698,458 (60.3%)
Individual	2,260,530 (14.1%)	2,017,579 (18.2%)
Medicaid	1,083,167 (6.8%)	978,177 (8.8%)
Medicare	2,567,666 (16.1%)	3,313,068 (29.8%)
Any fatal or non-fatal self-harm		
# visits included in analysis	15,249,031 (95.4%)	10,551,857 (95.0%)
# visits with 90-day event[b]	98,089 (0.64%)	34,764 (0.33%)
Suicide death		
# visits included in analysis	13,981,418 (87.5%)	9,714,817 (87.5%)
# visits with 90-day event[b]	3,199 (0.023%)	1,510 (0.016%)

Table 1: Characteristics of the patient visits included in the study, summarized by visit type (mental health specialty or general medical).
[a]Patients may have multiple types of insurance.
[b]Percentage calculated using only visits included in the analysis.

	Mental health $n = 15,986,946$	General medical $n = 11,104,580$
	n (%)	n (%)
Depression diagnosis		
0–3 months	8,515,731 (53.3%)	2,757,582 (24.8%)
4–12	7,443,083 (46.6%)	3,815,869 (34.4%)
13–60 months	7,884,537 (49.3%)	4,964,064 (44.7%)
Anxiety diagnosis		
0–3 months	8,041,059 (50.3%)	2,525,512 (22.7%)
4–12 months	7,110,947 (44.5%)	3,242,979 (29.2%)
13–60 months	7,555,451 (47.3%)	4,478,364 (40.3%)
Antidepressant fill		
0–3 months	7,967,051 (49.8%)	4,013,619 (36.1%)
4–12 months	7,731,060 (48.4%)	4,531,589 (40.8%)
13–60 months	7,993,183 (50.0%)	5,333,794 (48.0%)
Benzodiazepine fill		
0–3 months	3,888,653 (24.3%)	1,969,251 (17.7%)
4–12 months	4,229,825 (26.5%)	2,393,763 (21.6%)
13–60 months	5,421,777 (33.9%)	3,552,294 (32.0%)
Inpatient MH encounter		
0–3 months	1,112,531 (7.0%)	587,133 (5.3%)
4–12 months	1,309,140 (8.2%)	762,549 (6.9%)
13–60 months	2,377,345 (14.9%)	1,629,003 (14.7%)
Emergency department MH encounter		
0–3 months	1,746,647 (10.9%)	1,004,143 (9.0%)
4–12 months	2,099,814 (13.1%)	1,255,128 (11.3%)
13–60 months	3,548,115 (22.2%)	2,337,818 (21.1%)
Prior self-harm		
0–3 months	199,478 (1.2%)	41,295 (0.37%)
4–12 months	201,901 (1.3%)	55,047 (0.50%)
13–60 months	348,536 (2.2%)	130,139 (1.2%)
PHQ 9th item response 2 or 3		
0–3 months	546,085 (3.4%)	81,294 (0.73%)
4–12 months	489,087 (3.1%)	110,727 (1.0%)
13–60 months	398,428 (2.2%)	147,768 (1.3%)

Table 2: Summary of selected temporal predictors for patient visits included in the study. MH: mental health.

	Mental health visits		General medical visits	
	Any self-harm	Suicide death	Any self-harm	Suicide death
AUC	0.850	0.807	0.829	0.815
Sensitivity (%)				
90th percentile	60.2	50.0	59.4	51.0
95th percentile	45.7	35.1	47.2	38.4
99th percentile	19.0	10.9	23.5	18.9
PPV (%)				
90th percentile	3.8	0.10	1.9	0.08
95th percentile	5.7	0.14	3.0	0.11
99th percentile	11.6	0.21	7.6	0.28

Table 3: Performance of any self-harm (fatal and non-fatal) and suicide death prediction models including predictors from all time periods.

to AUC, sensitivity, and PPV, with the latter two measures evaluated using the 95th risk score percentile cut-point. Additional results for sensitivity and PPV at other cut-points are given in the Supplementary Material.

In Figure 2, we show the variable importance results for AUC predictiveness. Focusing first on the top left panel of Figure 2, we observe that the most recent predictors, capturing information from 0–3 months prior to the visit, show strong signal of importance for predicting any self-harm following a mental health visit. Compared to a model using predictors from months 0–60, the model using only months 4–60 shows substantially worse risk discrimination, with a decrease in AUC from 0.850 to 0.776 (VIM = 0.075, 95% CI 0.063–0.087). Likewise, removing the month 0–3 predictors from a model using months 0–12 results in a drop in AUC from 0.846 to 0.764 (VIM = 0.082, 95% CI 0.070–0.094). Predictors from 13–60 months appear somewhat less important: Comparing the 0–60 month model to the 0–12 month model corresponds to a decrease in AUC from 0.850 to 0.830 (VIM = 0.021, 95% CI 0.011–0.031), and comparing the 4–60 month model to 4–12 months corresponds to a decrease in AUC from 0.793 to 0.764 (VIM = 0.029, 95% CI 0.015–0.043).

We observe similar patterns for sensitivity (top left panel of Figure 3) and PPV (top left panel of Figure 4) using the 95th risk score percentile cut-point. Removing months 0–3 from a larger prediction model results in a drop in sensitivity from 45% to between 28% and 30%, corresponding to a VIM value between 15% and 17%. Removing months 13–60, conversely, decreases sensitivity by only 6%. There is strong evidence of non-zero importance for the variable groups in question for each of these model comparisons. For PPV, we observe a drop from between 5.6% and 5.7%

to between 3.8% and 4.0% when removing predictors from months 0–3, while the decrease from removal of months 13–60 is only 0.4%.

The bottom left panels of Figures 2–4 show the results for predicting the risk of any self-harm after a general medical visit. By and large, the results mirror those seen in the mental health setting, particularly for AUC and sensitivity. For PPV, the magnitude of estimated importance is lower for all model comparisons and is near zero for the month 13–60 variables. The smaller magnitude of estimated importance matches the lower event rate for any self-harm after a general medical visit versus a mental health specialty visit.

The right columns of Figures 2–4 show the estimated variable importance for predicting the risk of suicide death. Compared to the inferential results for predicting any self-harm, there is substantially more uncertainty in the suicide death analyses due to the smaller number of fatal self-harm events observed in the data set. The overall pattern of variable importance is similar between mental health and general medical visits. For AUC variable importance, the estimated importance of months 0–3 is around 0.05, corresponding to a decrease in AUC from 0.81 to 0.76, and the estimated importance of months 13–60 is close to zero. For sensitivity, predictors from months 0–3 have fairly large estimated importance relative to months 0–12 in both mental health (VIM = 15.3%, 95% CI 3.3%–27.2%) and general medical (VIM = 21.6%, 95% CI 9.3%–33.9%) settings. Month 0–3 predictors also demonstrate substantial importance relative to months 0–60 in the general medical setting (VIM = 17.8%, 95% CI 5.4%–30.1%). The ranking of variable importance estimates in terms of PPV is similar for predicting suicide death as for predicting any self-harm, although the confidence intervals are wide due to the low event rate and small absolute number of events.

In the Supplementary Material, we present results for sensitivity and PPV at cut-points based on the 90th and 99th percentiles of estimated risk; the overall patterns mimic those seen in Figures 3 and 4. The magnitude of the estimated VIMs varies by the percentile of risk score used as a cut-point. For sensitivity, for example, using a lower cut-point results in both higher sensitivity for all models (see Table 3) and larger VIM values, i.e., larger absolute differences between models. The opposite pattern is observed for PPV variable importance because PPV decreases as more visits are classified as high-risk.

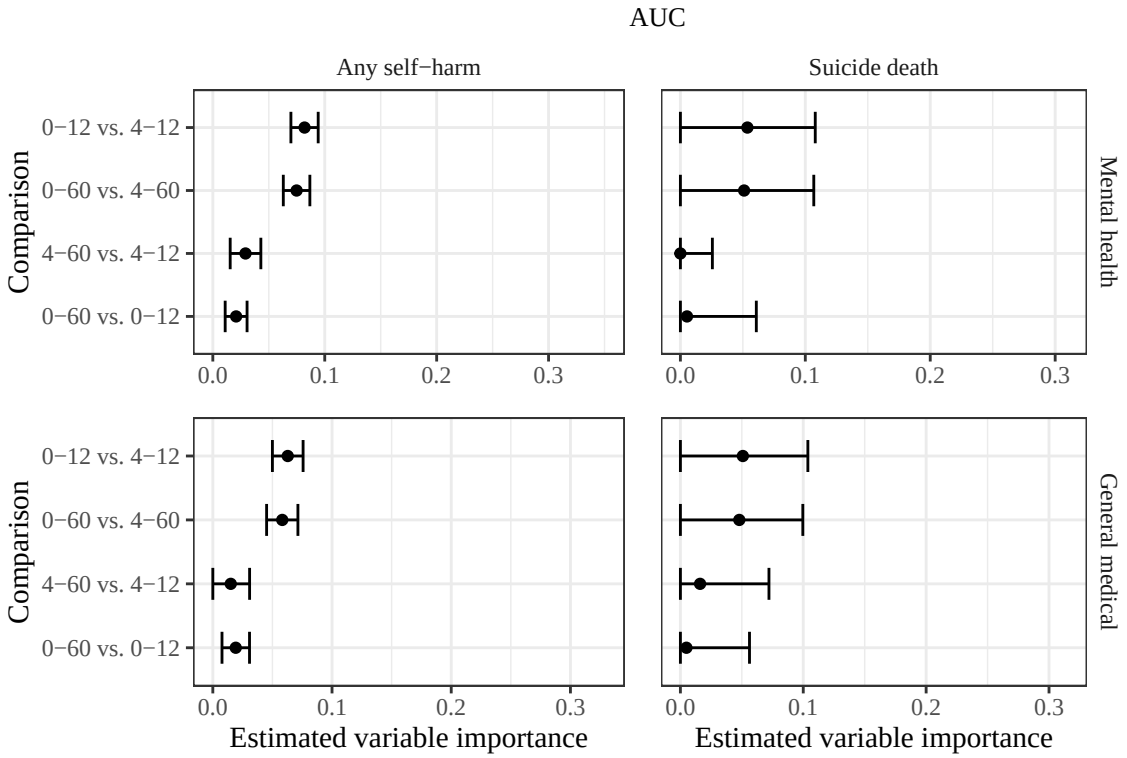

Figure 2: Estimated variable importance for temporal predictor groups in terms of **AUC**. Note the different x-axis scales for each outcome-setting pair, which are based on the estimated maximum possible variable importance (see Supplementary Material for details).

DISCUSSION

Using a sample of over 27 million visits made by patients across seven health systems, we assessed the importance of temporally grouped sets of variables for predicting the 90-day risk of any self-harm (fatal and non-fatal) and of suicide death. Using several predictiveness measures, we found consistent evidence that mental health–specific features corresponding to the most recent three months prior to the visit were highly important for predicting the risk of any self-harm. We found slightly weaker evidence for the importance of these features in predicting the risk of suicide death, although there was substantially larger uncertainty in evaluating the suicide death prediction models due to the smaller number of events. For prediction of any self-harm following a mental health specialty visit, removing predictors from the most recent three months resulted in a drop in AUC from 0.85 to 0.78, corresponding to a loss of nearly 20% of the discriminative potential of the model relative to the AUC of a null model (0.5). Features capturing patient information from one to five years prior to the prediction instance appear less important than more recent features.

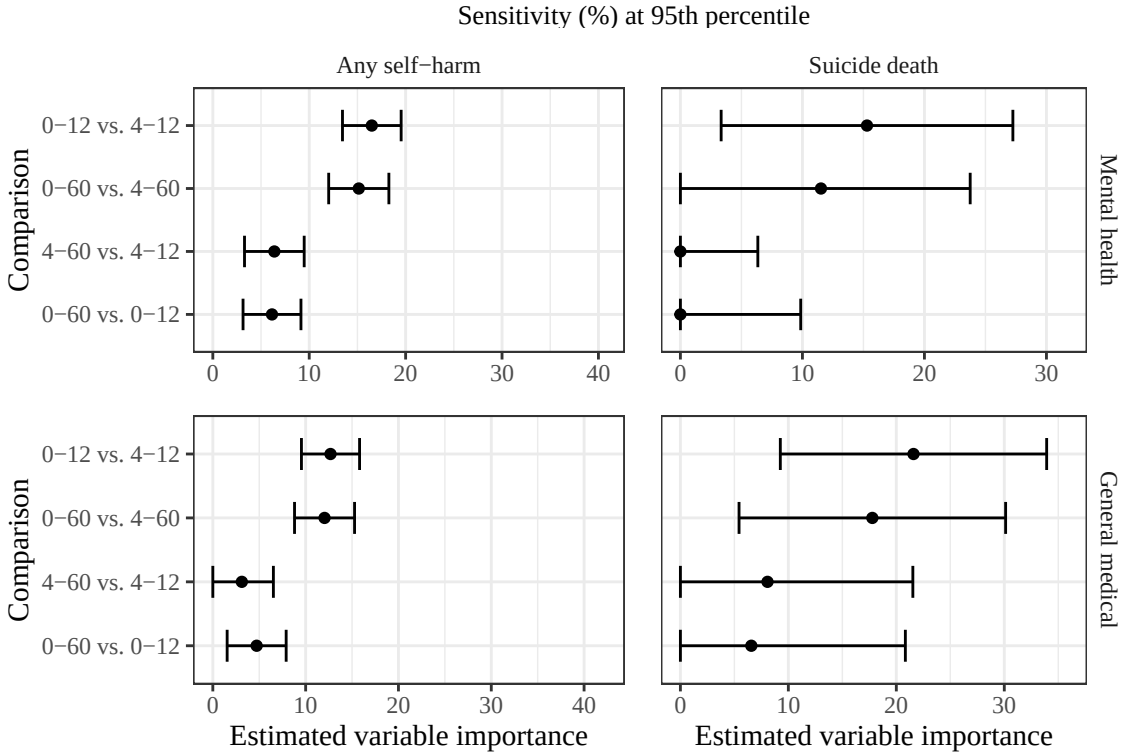

Figure 3: Estimated variable importance for temporal predictor groups in terms of **sensitivity at the 95th percentile of risk scores**. Note the different x-axis scales for each outcome-setting pair, which are based on the estimated maximum possible variable importance (see Supplementary Material for details).

One concern motivating this study was that complete information on recent predictors may not be available in real-time to include in risk calculations. This may be due to delays in processing health insurance claims: Information that may only be available in claims data (such as prescription fills or encounters with providers external to the health system) will not be immediately reflected in the health system data used to generate risk predictions at the time of the index visit. The impact of this lag cannot be easily examined in prediction modeling studies with retrospective cohort data, since it is challenging to determine what data would have been available at the time of an encounter. Thus, we used variable importance analyses to examine the predictive contribution of risk factors in the three months preceding a visit (a time period after which most claims data would be available). We found these recent predictors to be highly important, indicating that a model excluding this information would not as accurately identify patients at higher risk of self-harm following a visit. This result suggests that realizing good real-time predictive performance of self-harm prediction models would be unlikely for claims-only settings where recent data are not completely available at the time a prediction is made. In a health system with access to clinical and

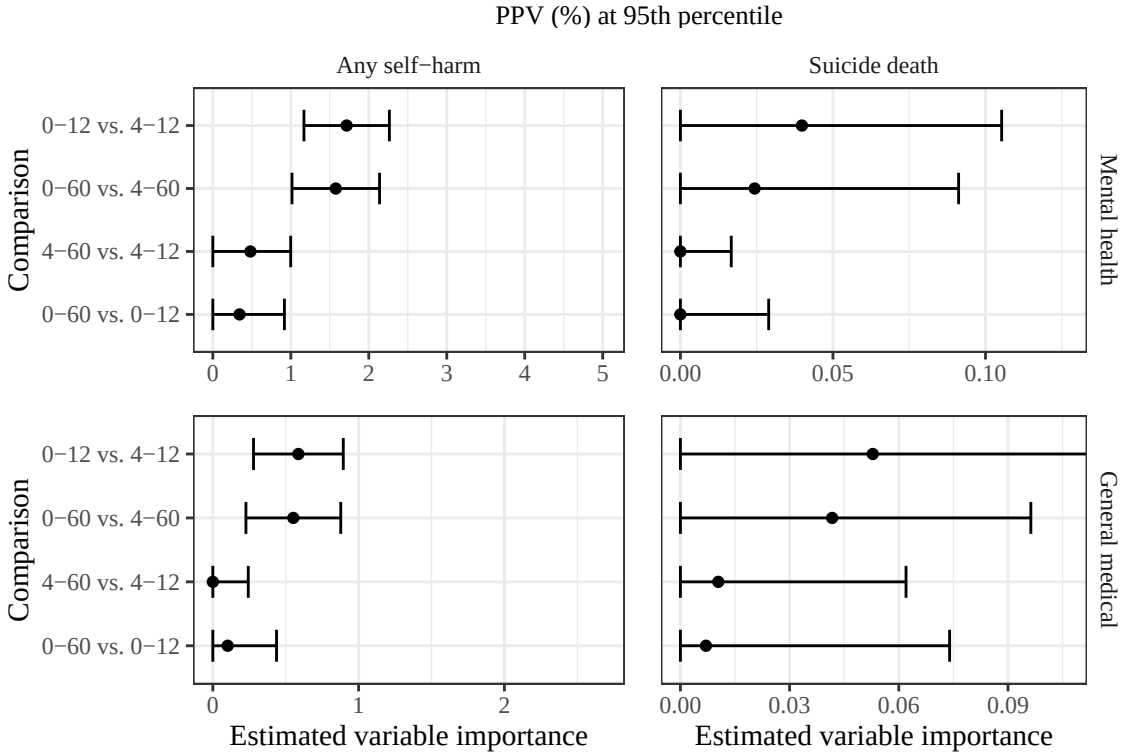

Figure 4: Estimated variable importance for temporal predictor groups in terms of **PPV at the 95th percentile** of risk scores. Note the different x-axis scales for each outcome-setting pair, which are based on the estimated maximum possible variable importance (see Supplementary Material for details).

claims data, we recommend prospectively monitoring availability of risk factors and performance of models with real-time data to quantify the impact of delayed data availability on identification of high-risk visits.

A second concern motivating this study was that patients without long-term, sustained insurance coverage would not have complete information on self-harm risk factors going back five years. Stable insurance coverage not only influences access to affordable health care but is also related to other social determinants of health including employment security and financial strain. As such, implementing a self-harm prediction model with differential capture of predictors could exacerbate existing health disparities, a concern frequently raised about the use of machine learning and artificial intelligence in clinical settings [17, 36–38]. This study found that incorporating predictors preceding an encounter by over a year provided only a small improvement in predictive performance. Thus, health systems could maintain strong overall risk identification without undercapturing variables for more recent enrollees by implementing a prediction model that excludes more distant predictors.

Variable importance analyses can provide a valuable tool for the development and implementation of prediction models in a variety of scenarios. For example, when transporting prediction models between settings, there may be concern that certain data elements are unavailable or incompletely captured in the new setting. If variables corresponding to such data elements are found to be important, investigators must carefully analyze the potential impacts on model performance. Assessing importance of predictors that are expensive or intrusive to collect can provide guidance on which variables or variable groups to prioritize for collection. Variable importance analyses could also be used to investigate concerns about equity in clinical prediction models, e.g., by quantifying the predictive contribution of variables expected to be most affected by structural racism and other institutionalized systems of disadvantage.

There are several opportunities for future research to extend the work of this study. First, this analysis included data from integrated health systems with access to both clinical records and health insurance claims data. The importance of clinical predictors from different time periods may vary for health systems with limited access to external claims data. Second, we examined variable importance for 90-day self-harm outcomes. The predictive value of risk factors measured at different time periods preceding the prediction instance may vary for different event horizons. Third, we did not evaluate variable importance within patient subgroups, such as those defined by type of insurance coverage or by race and ethnicity. Prior to clinical use of a prediction model, it is imperative to assess performance in subgroups of interest to ensure implementation does not lead to inequitable allocation of health care resources [17].

CONCLUSION

Self-harm risk prediction models often use predictors capturing up to five years of information on diagnoses, dispensed medications, and responses to the 9th item of the Patient Health Questionnaire. We found that the most recent three months of mental health–specific features were highly important for predicting the risk of non-fatal and fatal self-harm; removing these predictors resulted in a drop in AUC from 0.85 (all predictors) to 0.78. These findings suggest that rapid capture of recent data and integration into health records is crucial for predicting self-harm risk.

FUNDING

This research was supported by the National Institute of Mental Health (U19-MH099201, U19-MH121738, R01-MH125821) and by the National Science Foundation Graduate Research Fellowship Program (DGE-2140004).

AUTHOR CONTRIBUTIONS

C.J.W., B.D.W., and R.Y.C conceptualized and led the analysis. All authors contributed to the design and planning of study methodology. S.M.S., R.A.Z., G.E.S., R.C.R., B.K.A., Y.D., K.J.C., and F.L.L. acquired the data. C.J.W. and R.Y.C. analyzed the data. C.J.W., B.D.W., and R.Y.C. wrote the initial draft. All authors contributed revisions and approved the final version of the manuscript.

CONFLICTS OF INTEREST

K.J.C. has worked on grants awarded to Kaiser Permanente Southern California by Janssen Pharmaceuticals. S.M.S. has worked on grants awarded to Kaiser Permanente Washington Health Research Institute (KPWHRI) by Bristol Meyers Squibb and by Pfizer. She was also a co-investigator on grants awarded to KPWHRI from Syneos Health, who represented a consortium of pharmaceutical companies carrying out FDA-mandated studies regarding the safety of extended-release opioids. The other authors report there are no competing interests to declare.

DATA AVAILABILITY

The datasets generated and analyzed during this study are not publicly available because they contain detailed information from the electronic health records in the health systems participating in this study and are governed by HIPAA. Data are however available from the authors upon reasonable request, with permission of all health systems involved and fully executed data use agreement.

REFERENCES

[1] National Institute of Mental Health. Suicide. https://www.nimh.nih.gov/health/statistics/suicide#:~:text=The%20total%20age%2Dadjusted%20suicide,13.5%20per%20100%2C000%20in%202020, 2023. Accessed November 10, 2023.

[2] Centers for Disease Control and Prevention. Facts about suicide. https://www.cdc.gov/suicide/facts/index.html#:~:text=Suicide%20was%20responsible%20for%2048%2C183,one%20death%20every%2011%20minutes.&text=The%20number%20of%20people%20who,attempt%20suicide%20is%20even%20higher., 2023. Accessed March 25, 2024.

[3] Kurt Kroenke, Robert L Spitzer, and Janet BW Williams. The PHQ-9: validity of a brief depression severity measure. *Journal of General Internal Medicine*, 16(9):606–613, 2001.

[4] Douglas G Jacobs, Ross J Baldessarini, Yeates Conwell, Jan A Fawcett, Leslie Horton, Herbert Meltzer, Cynthia R Pfeffer, and Robert I Simon. Assessment and treatment of patients with suicidal behaviors. *APA Practice Guidelines*, 1:183, 2010.

[5] Yuval Barak-Corren, Victor M Castro, Solomon Javitt, Alison G Hoffnagle, Yael Dai, Roy H Perlis, Matthew K Nock, Jordan W Smoller, and Ben Y Reis. Predicting suicidal behavior from longitudinal electronic health records. *American Journal of Psychiatry*, 174(2):154–162, 2017.

[6] Ronald C Kessler, Murray B Stein, Maria V Petukhova, Paul Bliese, Robert M Bossarte, Evelyn J Bromet, Carol S Fullerton, Stephen E Gilman, Christopher Ivany, Lisa Lewandowski-Romps, A M Bell, J A Naifeh, M K Nock, B Y Reis, A J Rosellini, N A Sampson, A M Zaslavsky, and Ursano, R J, On behalf of the Army STARSS Collaborators. Predicting suicides after outpatient mental health visits in the Army Study to Assess Risk and Resilience in Servicemembers (Army STARRS). *Molecular Psychiatry*, 22(4):544–551, 2017.

[7] Colin G Walsh, Jessica D Ribeiro, and Joseph C Franklin. Predicting suicide attempts in adolescents with longitudinal clinical data and machine learning. *Journal of Child Psychology and Psychiatry*, 59(12):1261–1270, 2018.

[8] Gregory E Simon, Eric Johnson, Jean M Lawrence, Rebecca C Rossom, Brian Ahmedani, Frances L Lynch, Arne Beck, Beth Waitzfelder, Rebecca Ziebell, Robert B Penfold, and Susan M Shortreed. Predicting suicide attempts and suicide deaths following outpatient visits using electronic health records. *American Journal of Psychiatry*, 175(10):951–960, 2018.

[9] Jaimie L Gradus, Anthony J Rosellini, Erzsébet Horváth-Puhó, Amy E Street, Isaac Galatzer-Levy, Tammy Jiang, Timothy L Lash, and Henrik T Sørensen. Prediction of sex-specific suicide risk using machine learning and single-payer health care registry data from Denmark. *JAMA Psychiatry*, 77(1):25–34, 2020.

[10] Michael Sanderson, Andrew GM Bulloch, JianLi Wang, Kimberly G Williams, Tyler Williamson, and Scott B Patten. Predicting death by suicide following an emergency department visit for parasuicide with administrative health care system data and machine learning. *EClinicalMedicine*, 20, 2020.

[11] Le Zheng, Oliver Wang, Shiying Hao, Chengyin Ye, Modi Liu, Minjie Xia, Alex N Sabo, Liliana Markovic, Frank Stearns, Laura Kanov, Karl G Sylvester, Eric Widen, Doff B McElhinney, Wei

Zhang, Jiayu Liao, and Xuefeng B Ling. Development of an early-warning system for high-risk patients for suicide attempt using deep learning and electronic health records. *Translational Psychiatry*, 10(1):72, 2020.

[12] Fuchiang R Tsui, Lingyun Shi, Victor Ruiz, Neal D Ryan, Candice Biernesser, Satish Iyengar, Colin G Walsh, and David A Brent. Natural language processing and machine learning of electronic health records for prediction of first-time suicide attempts. *JAMIA Open*, 4(1): ooab011, 2021.

[13] Ilkin Bayramli, Victor Castro, Yuval Barak-Corren, Emily M Madsen, Matthew K Nock, Jordan W Smoller, and Ben Y Reis. Temporally informed random forests for suicide risk prediction. *Journal of the American Medical Informatics Association*, 29(1):62–71, 2022.

[14] Susan M Shortreed, Rod L Walker, Eric Johnson, Robert Wellman, Maricela Cruz, Rebecca Ziebell, R Yates Coley, Zimri S Yaseen, Sai Dharmarajan, Robert B Penfold, Brian K Ahmedani, Rebecca C Rossom, Arne Beck, Jennifer M Boggs, and Gregory E Simon. Complex modeling with detailed temporal predictors does not improve health records-based suicide risk prediction. *npj Digital Medicine*, 6(1):47, 2023.

[15] Gregory E Simon, Carolyn M Rutter, Do Peterson, Malia Oliver, Ursula Whiteside, Belinda Operskalski, and Evette J Ludman. Does response on the PHQ-9 depression questionnaire predict subsequent suicide attempt or suicide death? *Psychiatric Services*, 64(12):1195–1202, 2013.

[16] Qi Chen, Yanli Zhang-James, Eric J Barnett, Paul Lichtenstein, Jussi Jokinen, Brian M D'Onofrio, Stephen V Faraone, Henrik Larsson, and Seena Fazel. Predicting suicide attempt or suicide death following a visit to psychiatric specialty care: A machine learning study using Swedish national registry data. *PLoS Medicine*, 17(11):e1003416, 2020.

[17] R Yates Coley, Eric Johnson, Gregory E Simon, Maricela Cruz, and Susan M Shortreed. Racial/ethnic disparities in the performance of prediction models for death by suicide after mental health visits. *JAMA Psychiatry*, 78(7):726–734, 2021.

[18] Ulrike Grömping. Variable importance assessment in regression: linear regression versus random forest. *The American Statistician*, 63(4):308–319, 2009.

[19] Ivan Díaz, Alan Hubbard, Anna Decker, and Mitchell Cohen. Variable importance and prediction methods for longitudinal problems with missing variables. *PLOS ONE*, 10(3):e0120031, 2015.

[20] Scott M Lundberg and Su-In Lee. A unified approach to interpreting model predictions. *Advances in Neural Information Processing Systems*, 30, 2017.

[21] Jing Lei, Max G'Sell, Alessandro Rinaldo, Ryan J Tibshirani, and Larry Wasserman. Distribution-free predictive inference for regression. *Journal of the American Statistical Association*, 113(523):1094–1111, 2018.

[22] Luqin Gan, Lili Zheng, and Genevera I Allen. Inference for interpretable machine learning: Fast, model-agnostic confidence intervals for feature importance. *arXiv preprint arXiv:2206.02088*, 2022.

[23] Brian D Williamson, Peter B Gilbert, Noah R Simon, and Marco Carone. A general framework for inference on algorithm-agnostic variable importance. *Journal of the American Statistical Association*, 118(543):1645–1658, 2023.

[24] Brian D Williamson, Craig A Magaret, Peter B Gilbert, Sohail Nizam, Courtney Simmons, and David Benkeser. Super LeArner Prediction of NAb Panels (SLAPNAP): a containerized tool for predicting combination monoclonal broadly neutralizing antibody sensitivity. *Bioinformatics*, 37(22):4187–4192, 2021.

[25] Tyler R Ross, Daniel Ng, Jeffrey S Brown, Roy Pardee, Mark C Hornbrook, Gene Hart, and John F Steiner. The HMO research network virtual data warehouse: a public data model to support collaboration. *Egems*, 2(1), 2014.

[26] Shelly S Bakst, Tali Braun, Inbar Zucker, Ziva Amitai, and Tamy Shohat. The accuracy of suicide statistics: are true suicide deaths misclassified? *Social Psychiatry and Psychiatric Epidemiology*, 51:115–123, 2016.

[27] Kenneth L Cox, Matthew K Nock, Quinn M Biggs, Jennifer Bornemann, Lisa J Colpe, Catherine L Dempsey, Steven G Heeringa, James E McCarroll, Tsz Hin Ng, Michael Schoenbaum, Robert J Ursano, Bailey G Zhang, and David M Benedek. An examination of potential misclassification of army suicides: results from the army study to assess risk and resilience in servicemembers. *Suicide and Life-Threatening Behavior*, 47(3):257–265, 2017.

[28] Rebecca C Rossom, Karen J Coleman, Brian K Ahmedani, Arne Beck, Eric Johnson, Malia Oliver, and Greg E Simon. Suicidal ideation reported on the PHQ9 and risk of suicidal behavior across age groups. *Journal of Affective Disorders*, 215:77–84, 2017.

[29] Robert B Penfold, Ursula Whiteside, Eric E Johnson, Christine C Stewart, Malia M Oliver, Susan M Shortreed, Arne Beck, Karen J Coleman, Rebecca C Rossom, Jean M Lawrence, and Gregory E Simon. Utility of item 9 of the patient health questionnaire in the prospective identification of adolescents at risk of suicide attempt. *Suicide and Life-Threatening Behavior*, 51(5):854–863, 2021.

[30] James A Hanley and Barbara J McNeil. The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology*, 143(1):29–36, 1982.

[31] Charles J Wolock, Peter B Gilbert, Noah Simon, and Marco Carone. Nonparametric variable importance for time-to-event outcomes with application to prediction of HIV infection. *arXiv preprint arXiv:2311.12726*, 2023.

[32] Jerome Friedman, Trevor Hastie, and Rob Tibshirani. Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software*, 33(1):1, 2010.

[33] Trevor Hastie, Robert Tibshirani, Jerome H Friedman, and Jerome H Friedman. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, volume 2. Springer, 2009.

[34] Ben Dai, Xiaotong Shen, and Wei Pan. Significance tests of feature relevance for a black-box learner. *IEEE Transactions on Neural Networks and Learning Systems*, 2022. ISSN 21622388. doi: 10.1109/TNNLS.2022.3185742.

[35] Bradley Efron. Bootstrap Methods: Another Look at the Jackknife. *The Annals of Statistics*, 7(1):1 – 26, 1979. doi: 10.1214/aos/1176344552. URL `https://doi.org/10.1214/aos/1176344552`.

[36] Alvin Rajkomar, Michaela Hardt, Michael D Howell, Greg Corrado, and Marshall H Chin. Ensuring fairness in machine learning to advance health equity. *Annals of Internal Medicine*, 169(12):866–872, 2018.

[37] Ravi B Parikh, Stephanie Teeple, and Amol S Navathe. Addressing bias in artificial intelligence in health care. *JAMA*, 322(24):2377–2378, 2019.

[38] Milena A Gianfrancesco, Suzanne Tamang, Jinoos Yazdany, and Gabriela Schmajuk. Potential biases in machine learning algorithms using electronic health record data. *JAMA Internal Medicine*, 178(11):1544–1547, 2018.

Supplementary material

SETTING THE X-AXIS SCALE FOR FIGURES

For each combination of visit type and outcome, we calculated the maximum achievable variable importance using the performance of the full model including predictors from all time periods. For AUC, the maximum achievable importance is the difference between the AUC of the full model and 0.5 (the AUC of a null model). For sensitivity at a given cut-point, the maximum importance is the difference between the sensitivity of the full model and the proportion of visits flagged using that cut-point. For PPV, the maximum importance is the difference between the PPV of the full model and the overall event rate. In each figure, the upper limit of the x-axis is set to the maximum achievable importance.

SUPPLEMENTARY TABLES AND FIGURES

Health system	Data start date	Last date with complete cause of death data[a]
HealthPartners	January 1, 2009	December 31, 2016
Henry Ford Health	December 1, 2012[b]	December 31, 2015
Kaiser Permanente Colorado	January 1, 2009	December 31, 2017
Kaiser Permanente Hawaii	January 1, 2009	December 31, 2016
Kaiser Permanente Northwest	January 1, 2009	December 31, 2016
Kaiser Permanente Southern California	January 1, 2009	December 31, 2016
Kaiser Permanente Washington	January 1, 2009	December 31, 2016

Table S1: Data availability dates for participating sites.
[a]The study sample includes visits up to September 30 of the year with complete capture of cause of death data to allow for 90 days follow-up after mental health visits. For example, visits through September 30, 2016 are included for health systems with cause of death data complete through December 31, 2016.
[b]Only visits that occurred after the implementation of a new electronic health records system at Henry Ford were included in the sample.

Base predictors	
Age	Sex
Race	Hispanic ethnicity
Medicaid coverage at visit	Commercial insurance coverage at visit
Private pay insurance at visit	State-subsidized insurance at visit
Self-funded insurance at visit	Medicare insurance at visit
Other insurance at visit	High-deductible insurance at visit
Median household income <$25k at visit (census-based)	Median household income <$40k at visit (census-based)
Neighborhood <25% college educated at visit (census-based)	
Temporal predictors[a]	
Depression diagnosis	Anxiety diagnosis
Bipolar diagnosis	Schizophrenia diagnosis
Other psychological disorder diagnosis	Dementia diagnosis
ADD diagnosis	ASD diagnosis
Personality disorder diagnosis	Alcohol use disorder diagnosis
Drug use disorder diagnosis	PTSD diagnosis
Eating disorder diagnosis	Traumatic brain injury diagnosis
Antidepressant prescription fill	Benzodiazepine prescription fill
Hypnotic prescription fill	Second generation antipsychotic prescription fill
Inpatient encounter with MH diagnosis	Outpatient MH specialty visit
Emergency/urgent care encounter MH diagnosis	Any self-inflicted injury/poisoning
Self-inflicted lacerative violent injury	Other self-inflicted violent injury
Any injury/poisoning diagnosis	Natal delivery diagnosis
Modal PHQ9 9th item response	Maximum PHQ9 9th item response
Number of PHQ9 9th item responses	

Table S2: Overview of variables included in prediction models. MH: mental health.
[a]Each temporal predictor is a binary indicator of presence/absence in 0–3 months, 3 months – 1 year, or 1–5 years prior to the index visit.

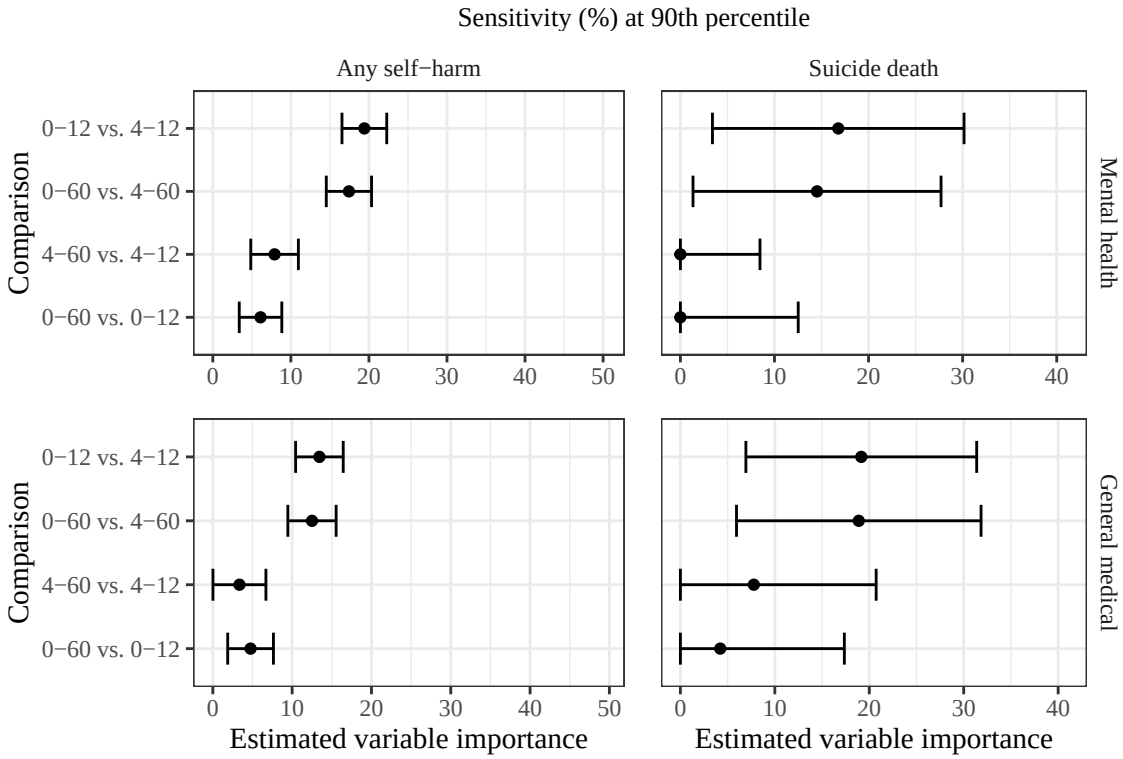

Figure S1: Estimated variable importance for temporal predictor groups in terms of **sensitivity at the 90th percentile** of risk scores. Note the different x-axis scales for each outcome-setting pair, which are based on the estimated maximum possible variable importance (see Supplementary Material for details).

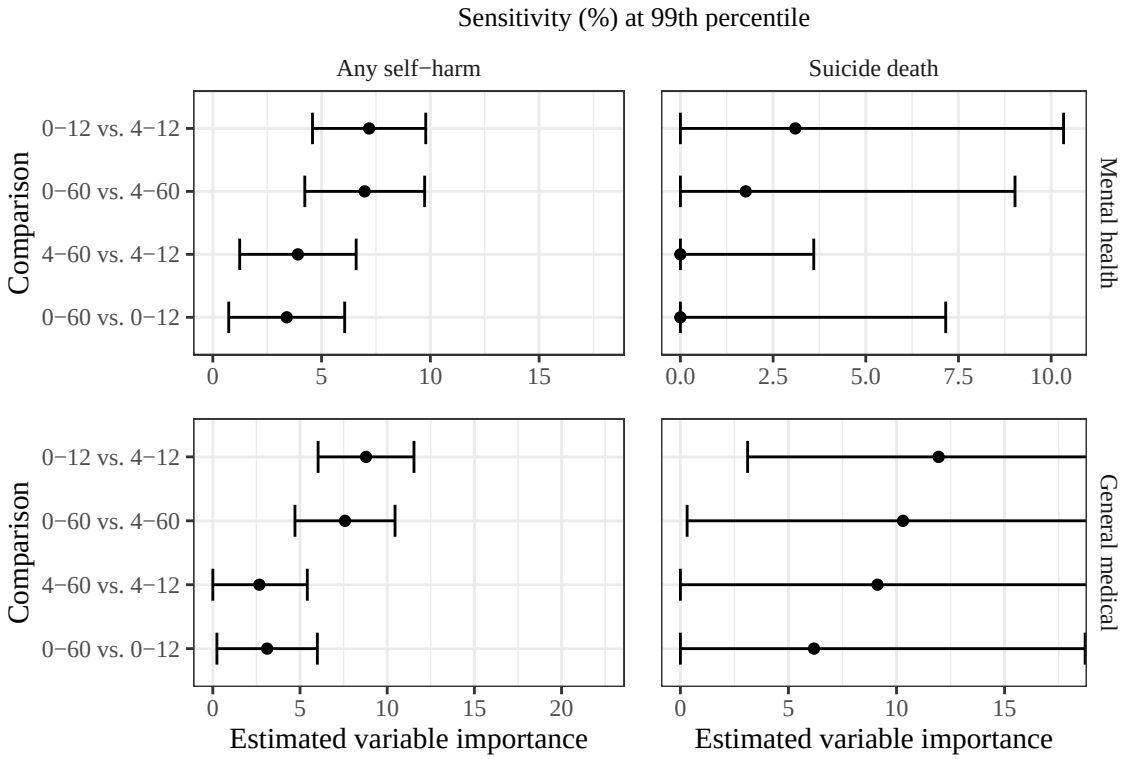

Figure S2: Estimated variable importance for temporal predictor groups in terms of **sensitivity at the 99th percentile** of risk scores. Note the different x-axis scales for each outcome-setting pair, which are based on the estimated maximum possible variable importance (see Supplementary Material for details).

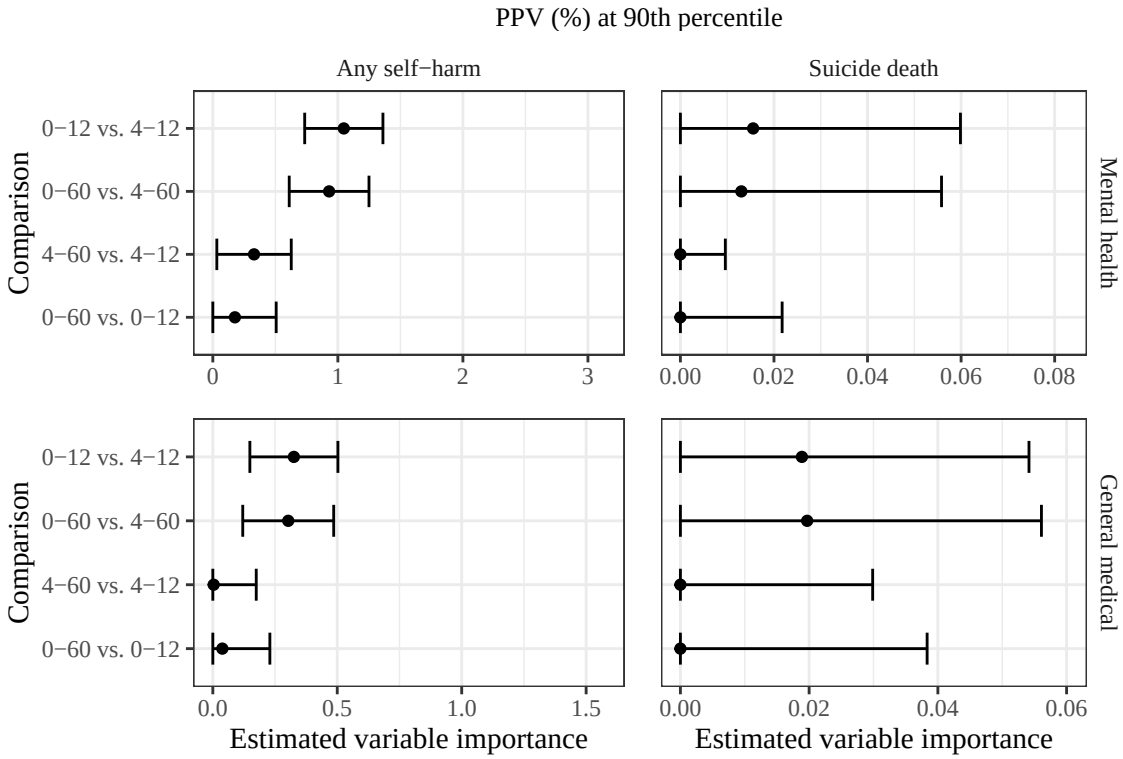

Figure S3: Estimated variable importance for temporal predictor groups in terms of **PPV at the 90th percentile** of risk scores. Note the different x-axis scales for each outcome-setting pair, which are based on the estimated maximum possible variable importance (see Supplementary Material for details).

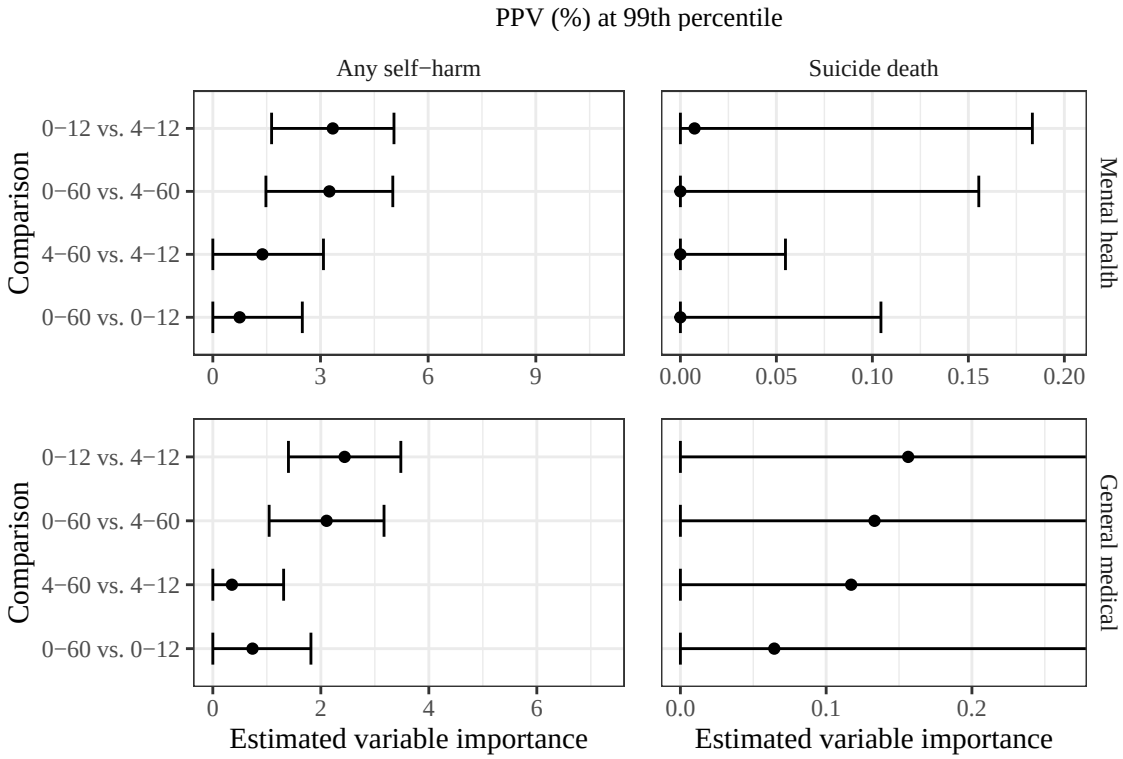

Figure S4: Estimated variable importance for temporal predictor groups in terms of **PPV at the 99th percentile** of risk scores. Note the different x-axis scales for each outcome-setting pair, which are based on the estimated maximum possible variable importance (see Supplementary Material for details).