Article published online: 2021-08-18

Evaluation of Electronic Health Record-Based Suicide Risk Prediction Models on Contemporary Data

Rod L. Walker[1] Susan M. Shortreed[1] Rebecca A. Ziebell[1] Eric Johnson[1] Jennifer M. Boggs[2] Frances L. Lynch[3] Yihe G. Daida[4] Brian K. Ahmedani[5] Rebecca Rossom[6] Karen J. Coleman[7] Gregory E. Simon[1]

[1] Kaiser Permanente Washington Health Research Institute, Seattle, Washington, United States
[2] Kaiser Permanente Colorado, Institute for Health Research, Aurora, Colorado, United States
[3] Kaiser Permanente Northwest, Center for Health Research, Portland, Oregon, United States
[4] Kaiser Permanente Hawaii, Center for Integrated Health Care Research, Honolulu, Hawaii, United States
[5] Henry Ford Health System, Center for Health Policy & Health Services Research, Detroit, Michigan, United States
[6] Department of Research, HealthPartners Institute, Minneapolis, Minnesota, United States
[7] Department of Research and Evaluation, Kaiser Permanente Southern California, Pasadena, California, United States

Address for correspondence Rod L. Walker, MS, Kaiser Permanente Washington Health Research Institute, 1730 Minor Avenue, Suite 1600, Seattle, WA 98101, United States (e-mail: Rod.L.Walker@kp.org).

Appl Clin Inform 2021;12:778–787.

Abstract

Background Suicide risk prediction models have been developed by using information from patients' electronic health records (EHR), but the time elapsed between model development and health system implementation is often substantial. Temporal changes in health systems and EHR coding practices necessitate the evaluation of such models in more contemporary data.

Objectives A set of published suicide risk prediction models developed by using EHR data from 2009 to 2015 across seven health systems reported c-statistics of 0.85 for suicide attempt and 0.83 to 0.86 for suicide death. Our objective was to evaluate these models' performance with contemporary data (2014–2017) from these systems.

Methods We evaluated performance using mental health visits (6,832,439 to mental health specialty providers and 3,987,078 to general medical providers) from 2014 to 2017 made by 1,799,765 patients aged 13+ across the health systems. No visits in our evaluation were used in the previous model development. Outcomes were suicide attempt (health system records) and suicide death (state death certificates) within 90 days following a visit. We assessed calibration and computed c-statistics with 95% confidence intervals (CI) and cut-point specific estimates of sensitivity, specificity, and positive/negative predictive value.

Results Models were well calibrated; 46% of suicide attempts and 35% of suicide deaths in the mental health specialty sample were preceded by a visit (within 90 days) with a risk score in the top 5%. In the general medical sample, 53% of attempts and 35% of deaths were preceded by such a visit. Among these two samples, respectively, c-statistics were 0.862 (95% CI: 0.860–0.864) and 0.864 (95% CI: 0.860–0.869) for suicide attempt, and 0.806 (95% CI: 0.790–0.822) and 0.804 (95% CI: 0.782–0.829) for suicide death.

Conclusion Performance of the risk prediction models in this contemporary sample was similar to historical estimates for suicide attempt but modestly lower for suicide death. These published models can inform clinical practice and patient care today.

Keywords
- suicide
- mental health
- electronic health records
- statistics
- health care systems

received
February 16, 2021
accepted after revision
July 1, 2021

Georg Thieme Verlag KG,
Rüdigerstraße 14,
70469 Stuttgart, Germany

DOI https://doi.org/10.1055/s-0041-1733908.
ISSN 1869-0327.

Background and Significance

Identification of individuals at increased risk for suicide is an ongoing challenge in the United States. Traditional clinical tools such as depression questionnaires, specifically item 9 of the Patient Health Questionnaire (PHQ-9), can identify patients at high risk for suicide attempt or death but exhibit only moderate sensitivity.[1] More recent research, including our own, has attempted to improve the ability to predict suicidal behavior by capitalizing on large health system datasets and utilizing machine learning methods to develop models that draw upon the growing wealth of information captured in electronic health records (EHR).[2–6] While some EHR-based models may appear to have sufficient discriminative ability for possible clinical use, there exist many important considerations when planning potential implementation into a health system.

One such consideration, which serves as the impetus for this manuscript, is the issue of temporality. The time elapsed between when a risk prediction model is developed and when it is implemented within a health system is often substantial, and temporal changes in health systems and EHR coding practices can be notable.[7–10] Such changes can call into question whether performance estimates generated at the model development stage accurately reflect the performance that will be observed at the time when a model is to be put into practice. As such, it is imperative to understand whether models developed on data during an older period are adequate or need to be refit or redeveloped with more recent data, a process that can be time and resource intensive.

Objectives

Our prior research collaboration pooled EHR data from 2009 to 2015 across seven health systems to develop models to predict risk of suicide attempt and suicide death following a mental health visit.[5] The resulting models demonstrated good discriminative performance based on the hold-out validation data at the time. C-statistics and 95% confidence intervals (CI) for the prediction of suicide attempt were 0.851 (95% CI: 0.848–0.853) for mental health specialty visits and 0.853 (95% CI: 0.849–0.857) for general medical visits (predominately made to primary care providers), while c-statistics for the prediction of suicide death were 0.861 (95% CI: 0.848, 0.875) and 0.833 (95% CI: 0.813–0.853), respectively. Numerous changes, however, have occurred since the original work: the shift from the ICD-9-CM (International Classification of Diseases, Ninth Revision, Clinical Modification) to ICD-10-CM (Tenth Revision) system for coding diagnoses; increased clinical use of standard depression questionnaires; health system implementation of suicide risk assessment protocols[11]; and shifts in health system membership. Such changes could impact performance of the published models; therefore, our goal in this current study was to validate those models by using EHR data from the same health systems in more recent years (2014–2017).

Methods

Setting

The seven health systems contributing to the previous and current work include HealthPartners (Minnesota); Henry Ford Health System (Michigan); and the Colorado, Hawaii, Northwest (Oregon), Southern California, and Washington regions of Kaiser Permanente. These integrated health care delivery systems provide both insurance coverage and comprehensive health care to patient-populations enrolled in their systems via individual or employer-sponsored insurance plans, Medicaid or Medicare, or other subsidized low-income insurance programs. The populations covered by these systems are representative of the systems' respective service areas, reflecting the demographic and socioeconomic characteristics of the covered geographic regions.

As part of their comprehensive health care delivery, each system provides both general medical and specialty mental health care. Further, each of these systems has established clinical recommendations to use the PHQ-9 at all visits for depression, including mental health specialty and primary care visits. However, the implementation of these guidelines have varied across the systems over time.[11] Additionally, all seven health systems participate as members of the Mental Health Research Network, with each maintaining a research data warehouse in line with the specifications model of the Health Care Systems Research Network's Virtual Data Warehouse.[12] Thus, each site has an electronic data resource for research that combines extensive information on the patient populations, including data on health plan enrollment, medical information captured in the EHR, claims, pharmacy dispensings, state mortality records, and neighborhood characteristics derived from census reporting. The responsible institutional review boards for each health system approved use of de-identified records data for the current research study.

Sample

Our original suicide risk prediction models were developed to predict suicide attempts following two types of health system encounters among health plan members aged 13 years and older: (1) mental health specialty visits defined as any outpatient clinic visit with a mental health specialty provider, and (2) general medical visits defined as any outpatient clinic visit with a nonmental health specialty provider (mostly primary care) at which a mental health or substance use diagnosis was recorded. Models were originally developed and validated by using such visits from each health system from January 1, 2009 to June 30, 2015, with predictions made at the visit level rather than the person level because a person's risk of suicide changes over time. Hereafter, we refer to visits used for the previous modeling work as the original sample.

For the current validation study, we collected a contemporary sample of mental health specialty and general medical visits that were not included in the previous model development and validation work. This contemporary sample included visits from January 1, 2014 to September 30,

2017 among people who were aged 13 years and older and enrolled in the health plan at the time of the visit. Our evaluation of suicide attempt models only used visits from the contemporary sample that occurred after the transition to ICD-10-CM (October 1, 2015–September 30, 2017). Our evaluation of suicide death models, however, used any visits from the contemporary sample that had cause-of-death information available. Coding of suicide deaths was not affected by the transition to ICD-10-CM but does rely on state mortality records, which are variably updated at each health system because each state releases death records at different intervals (available through 2016 for most sites in the contemporary sample). In the prior study, cause-of-death information was only available through 2013 at most sites, which is why we could include some visits from 2014 and the first half of 2015 in the contemporary sample (as they were not originally used to train or validate the risk models for suicide death).

In addition to the coding transition and temporal changes in health systems described earlier, data underlying the contemporary sample were also affected by technical changes (relative to the original sample) specific to research data infrastructure in these health systems, including different specialty codes used to include mental health specialty visits; narrower definition of substance use disorder used to include general medical visits; and more accurate health system enrollment information used to censor observations without complete follow-up data.

Outcomes

For each visit in the contemporary sample, we identified suicide attempts or suicide deaths within 90 days following the visit. Suicide attempts were ascertained by using EHR or insurance claim International Classification of Diseases, Tenth Revision, Clinical Modification (ICD-10-CM) diagnoses of injury or poisoning coded as either intentional self-harm or having undetermined intent. Suicide deaths were ascertained by using cause-of-death information obtained from state vital records; deaths with causes corresponding to ICD-10-CM diagnoses of intentional self-harm (codes X60-X84, Y87.0) or undetermined intent (codes Y10-Y34, Y87.2) were classified as suicides. Inclusion of instances with undetermined intent in the suicide attempt and suicide death outcomes significantly increases ascertainment of both, with a modest trade-off of specificity.[1] Suicide attempt ascertainment was censored at the time of a person's disenrollment from the health system because data regarding self-harm diagnoses would no longer be available after that time. Suicide death is ascertained regardless of enrollment status; therefore, no censoring was applied to that outcome.

Covariates

Model predictors were based on patient information gathered from health system records over the 5 years preceding each visit. These predictors included patient demographics such as age, sex, race/ethnicity, insurance type, and census-derived education and income measures; mental health and substance use disorder diagnoses (current and past), as well as more general medical diagnoses of comorbidity (components of Charlson Comorbidity Index[13]); previous suicide attempts or other injuries or poisonings; previous hospitalizations or emergency department encounters for mental health care; mental health medication dispensings; and PHQ-9 scores (total and item 9). Timing of diagnoses, prescriptions, PHQ-9 scores, previous suicide attempts, and utilization events were represented by using different combinations of indicators for occurrence at index, in the past three months, in the past six months, in the past year, or in the past five years, depending on the predictor category. Full detail on the set of predictors considered during development of the original risk prediction models are provided in Appendix 9A of the previous publication[5] and at github.com/MHResearchNetwork/srpm-analytic.

Statistical Analysis

As this is a validation study of an existing set of published risk prediction models, we summarize how those models were previously developed. Briefly, they were developed using logistic regression with least absolute shrinkage and selection operator (LASSO) variable selection[14] based on a random training sample of visits, with final models calibrated by refitting logistic regression models to the training sample using only the variables selected by LASSO and estimating coefficients using generalized estimating equations.[15] The published performance was based on applying final models to a held-out validation set of visits. The variable selection, model calibration, and validation steps were repeated for each of the outcomes of interest (90-day suicide attempt and suicide death) and separately for the mental health specialty and general medical samples. Details of the final selected predictors and summary of results from that previous work are provided in that manuscript[5] and at github.com/MHResearchNetwork/srpm-model.

For this current validation study, we applied those previously developed risk prediction models to the contemporary sample and calculated the same performance metrics that were presented in the original paper. These included (1) comparing predicted risk to observed risk within predicted risk strata (with strata cut-points defined using the predictions from the previous paper's original training dataset); (2) computing c-statistics measuring the area under the receiver operating characteristic (ROC) curves; and (3) calculating sensitivity, specificity, positive predictive value (PPV), and negative predictive value (NPV) associated with specific predicted risk cut-points (as determined in the original paper). Confidence intervals for c-statistics were computed via bootstrap with 1,000 replications.

In our evaluation of suicide attempt models, we excluded visits that were censored due to health system disenrollment. However, as a sensitivity analysis, we incorporated inverse probability of censoring weights to assess if censoring affected performance estimates. To assess the impact that technical changes in research data infrastructure (described earlier) may have had on differences in prediction performance between the original and contemporary samples, we collected mental health visits spanning the same years as the

original sample (2009–2015) but reflecting the current research infrastructure and data specifications. We then applied the published suicide attempt and suicide death models to this data, computed performance metrics, and compared with the originally published performance.

Model predictions were computed and performance assessed by using R version 3.5.3[16] and the ROCR package.[17]

Results

The contemporary sample included 6,832,439 mental health specialty visits and 3,987,078 general medical visits among 1,799,765 unique patients (►Table 1). Relative to the mental health specialty sample, patients from the general medical sample were older (26 vs. 13% of patients were aged 65 +), less likely to belong to a racial or ethnic minority group (70 vs. 66% white and 23 vs. 28% Hispanic), and less likely to be insured by commercial insurance (57 vs. 69%). As expected, the PHQ-9 item 9 score was much more frequently assessed in the mental health specialty sample (48% of patients assessed at the visit or within the past year compared with 27% in the general medical sample). Likewise, suicide rates were higher following mental health specialty visits as compared with general medical visits (63.3 vs. 32.0 per 10,000 visits for 90-day suicide attempts and 1.7 vs. 1.2 per 10,000 visits for 90-day suicide deaths).

Table 1 Patient characteristics among visits included in the contemporary sample

Characteristic	Mental health specialty		General medical	
	n	%	*n*	%
Total visits	6,832,439		3,987,078	
Total people	859,543		1,439,127	
Female	4,405,519	64	2,543,035	64
Age				
13–17	626,541	9	176,875	4
18–29	1,225,903	18	540,823	14
30–44	1,767,273	26	830,412	21
45–64	2,311,333	34	1,393,998	35
65 or older	901,389	13	1,044,970	26
Race				
White	4,482,143	66	2,780,067	70
Asian	374,147	5	181,036	5
Black	625,306	9	326,115	8
Hawaiian/Pacific Islander	69,069	1	33,345	1
Native American	60,878	1	40,848	1
More than one or other	23,550	0	26,727	1
Not recorded	1,197,346	18	598,940	15
Ethnicity				
Hispanic	1,899,758	28	930,007	23
Insurance type				
Commercial group	4,714,442	69	2,283,033	57
Individual	1,127,044	17	762,452	19
Medicare	442,189	6	570,165	14
Medicaid	520,331	8	308,964	8
Other	28,433	0	62,464	2
PHQ-9 Item 9 score recorded				
At index visit	1,564,602	23	504,061	13
At index visit or any visit in past year	3,280,778	48	1,056,676	27
Length of enrollment prior to visit				
1 y or more	5,910,410	87	3,274,462	82
5 y or more	3,831,454	56	2,033,333	51

Abbreviation: PHQ-9, Patient Health Questionnaire-9.

Table 2 Comparison of the distributions of select predictors between the original sample and the current (contemporary) sample

Predictor	Mental health specialty		General medical	
	Original %	Current %	Original %	Current %
Diagnoses in prior 5 y[a]				
Depressive disorder	74	77	55	66
Anxiety disorder	62	80	46	71
Bipolar disorder	13	13	6	6
Schizophrenia spectrum disorder	4	4	2	2
Personality disorder	6	16	2	10
Alcohol use disorder	13	15	6	13
Medication dispensing in prior 3 mo				
Antidepressant	44	51	30	36
Benzodiazepine	24	23	17	17
Hypnotic	5	3	4	2
Second-generation antipsychotic	12	13	4	4
Mental health clinical encounter in prior 3 mo				
Inpatient stay	7	7	5	5
Emergency department	10	12	8	9
Outpatient	79	79	21	19
Suicide attempt in the prior year	2	2	1	1
PHQ-9 Item 9 score recorded in prior year[a]	48	20	27	11

Abbreviation: PHQ-9, Patient Health Questionnaire-9.
[a]Includes index visit.

►**Table 2** provides a comparison of the distributions of select predictors between the contemporary sample and the original sample. Some notable differences were that the contemporary sample tended to have higher prevalence of anxiety disorder diagnoses, personality disorder diagnoses, and greater use of antidepressant medications, as well as a greater proportion receiving PHQ-9 item 9 assessments. **Supplementary Material 1** provides a more extensive set of comparisons. Rates of suicide attempts and suicide deaths were comparable in the contemporary and original samples.

Evaluation of the 90-day suicide attempt prediction models (which was limited to visits occurring after transition to ICD-10-CM) utilized 4,073,012 mental health specialty visits and 2,327,499 general medical visits from the contemporary sample and identified 9,109 unique suicide attempts within 90 days of a visit. Reflected in these counts is the exclusion of 3.5% of visits due to disenrollment within 90 days. Evaluation of the 90-day suicide death prediction models (which was limited to visits with cause-of-death information available from state vital records) utilized 4,799,175 mental health specialty visits, and 2,773,976 general medical visits from the contemporary sample and identified 356 unique suicide deaths within 90 days of a visit.

►**Table 3** compares the predicted and observed risk for the 90-day suicide attempt and suicide death risk prediction models in the contemporary sample. The largest differences were seen in the highest risk strata. For example, the average predicted risk of suicide attempt for mental health specialty visits in the top strata was 17.4% compared with an observed risk of 12.0% (0.84 vs. 0.48% for suicide death). For general medical visits, these predicted and observed risk were 10.6 and 9.2%, respectively, for suicide attempt, and 0.50 versus 0.13% for suicide death. Predicted and observed risk for other strata, which contain the majorities of events in the sample, were very similar. In the mental health specialty sample, 46% of suicide attempts and 35% of suicide deaths were preceded by at least one visit (within 90 days) in the top three risk strata (i.e., above the 95th percentile of predicted risk) even though those strata only accounted for approximately 5% of visits. Similarly, in the general medical sample, 53% of suicide attempts and 35% of suicide deaths were preceded by at least one visit in those top three strata, yet those strata included only approximately 7% of visits.

►**Fig. 1** provides ROC curves illustrating overall discrimination for the 90-day suicide attempt and suicide death risk prediction models in the contemporary sample. Among mental health specialty and general medical visits, respectively, c-statistics were 0.862 (95% CI: 0.860–0.864) and 0.864 (95% CI: 0.860–0.869) for suicide attempt, and 0.806 (95% CI: 0.790–0.822) and 0.804 (95% CI: 0.782–0.829) for suicide death. ►**Table 4** presents sensitivity, specificity, PPV, and NPV for each of the models at specific cut-points. For the suicide attempt models, for example, using a 99th percentile cut-point yielded a sensitivity of 22.2% and PPV of 9.9% for the mental health specialty visits and a sensitivity of 29.2% and

Table 3 Assessment of model calibration in contemporary cohort by comparing predicted risk to actual risk in predefined risk strata

Mental health specialty	**90-day suicide attempt**			**90-day suicide death**		
Risk score percentile strata	Predicted risk[a] (%)	Observed risk[b] (%)	% of all attempts[c]	Predicted risk[a] (%)	Observed risk[b] (%)	% of all deaths[c]
≥99.5th	17.4	12.0	15	0.84	0.48	12
99–99.5th	9.0	7.2	7	0.36	0.35	9
95–99th	3.9	3.5	26	0.15	0.07	14
90–95th	1.6	1.7	15	0.07	0.05	13
75–90th	0.8	0.9	20	0.03	0.02	20
50–75th	0.3	0.3	11	0.01	0.01	17
<50th	0.1	0.1	6	0.00	0.00	15
General medical	**90-day suicide attempt**			**90-day suicide death**		
Risk score percentile strata	Predicted risk[a] (%)	Observed risk[b] (%)	% of All attempts[c]	Predicted risk[a] (%)	Observed risk[a] (%)	% of all deaths[c]
≥99.5th	10.6	9.2	21	0.50	0.13	7
99–99.5th	3.9	3.7	8	0.20	0.05	3
95–99th	1.5	1.7	28	0.08	0.07	28
90–95th	0.6	0.7	13	0.04	0.03	13
75–90th	0.3	0.3	14	0.02	0.02	24
50–75th	0.1	0.1	9	0.01	0.01	13
<50th	0.1	0.0	6	0.00	0.00	12

[a]Average predicted risk among visits in this stratum.
[b]Proportion of visits in this stratum with an event within 90 days.
[c]Among all visits in the entire sample with an event within 90 days, the percentage of such visits within this stratum.
Note: Cut-points used to define risk score percentile strata are based on the distribution of risk predictions as observed in the original model development sample.

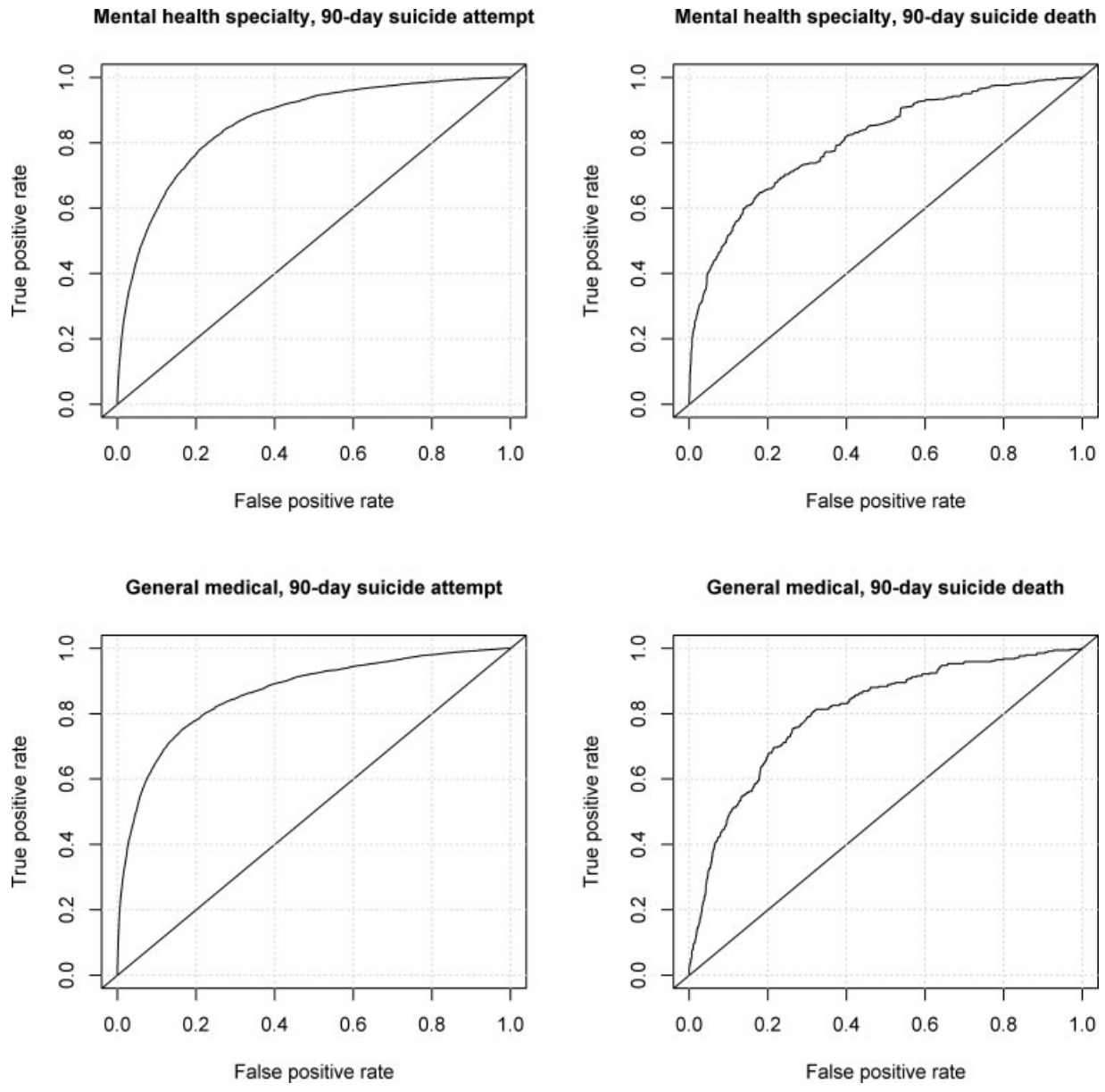

Fig. 1 Receiver operating characteristic curves illustrating model discrimination.

PPV of 6.5% for the general medical sample visits. Using a 95th percentile cut-point, these same metrics were 48.0 and 5.0%, and 57.4 and 2.7%, respectively. For the suicide death models, sensitivity tended to be lower than was observed for suicide attempt models for both the mental health specialty and the general medical visits. PPV was extremely low, as expected, due to the rarity of suicide death.

Sensitivity analyses incorporating weights to account for incomplete 90-day follow-up of suicide attempts yielded weighted c-statistics of 0.862 (95% CI: 0.860–0.864) and 0.864 (95% CI: 0.860–0.869) for the mental health specialty and general medical visits in the contemporary sample. Additionally, assessing performance of the models applied to visits reflecting the current data infrastructure but spanning the years of the original sample resulted in c-statistics that were near identical to the original published performance, thus suggesting that technical changes to data infrastructure (new specification of mental health specialty visits, narrower definitions of substance use disorder, and improved identification of insurance disenrollment) had no meaningful impact on model performance.

Discussion

We evaluated previously developed suicide risk models[5] on a contemporary sample of approximately 10 million mental health specialty and general medical visits among nearly 1 million patients across seven health systems. As the previously developed models had not been clinically deployed, we did not expect relationships between predictors and outcomes to have meaningfully changed.[18] Still, this evaluation was a necessary temporal validation[19] given other changes that occurred across health systems. We found that models were well calibrated in this contemporary sample despite

Table 4 Model performance characteristics at various cut-points of predicted risk[a]

Mental health specialty	90-day suicide attempt				90-day suicide death			
Risk score percentile cut-points	Sensitivity (%)	Specificity (%)	PPV (%)	NPV (%)	Sensitivity (%)	Specificity (%)	PPV (%)	NPV (%)
≥99th	22.2	98.7	9.9	99.5	21.0	99.1	0.42	100.0
≥95th	48.0	94.2	5.0	99.6	35.1	95.8	0.14	100.0
≥90th	62.8	88.7	3.4	99.7	48.6	91.4	0.10	100.0
≥75th	83.0	73.6	2.0	99.9	68.1	77.5	0.05	100.0
≥50th	94.2	50.1	1.2	99.9	85.4	52.6	0.03	100.0
General medical	**90-day suicide attempt**				**90-day suicide death**			
Risk score percentile cut-points	Sensitivity (%)	Specificity (%)	PPV (%)	NPV (%)	Sensitivity (%)	Specificity (%)	PPV (%)	NPV (%)
≥99th	29.2	98.7	6.5	99.8	9.9	98.6	0.09	100.0
≥95th	57.4	93.5	2.7	99.9	38.2	93.9	0.08	100.0
≥90th	70.4	87.4	1.8	99.9	51.3	88.5	0.05	100.0
≥75th	84.8	69.7	0.9	99.9	75.8	72.3	0.03	100.0
≥50th	94.3	40.8	0.5	100.0	88.3	50.0	0.02	100.0

Abbreviations: NPV, negative predictive value; PPV, positive predictive value.
[a]Cut-points are based on the distribution of risk predictions as observed in the original model development sample.

temporal changes in clinical practice and informatics changes within the health systems that included a transition from ICD-9-CM to ICD-10-CM coding. While the models tended to overestimate risk for visits in the highest predicted risk strata (e.g., above the 99.5th percentile), visits in these strata did have the highest observed risk in the sample, thus suggesting that those thresholds do identify high risk visits. Models also demonstrated good discriminative ability in stratifying visits according to 90-day risk of suicide attempt or suicide death for both the mental health specialty and general medical visits. The c-statistics for suicide attempt models in this contemporary data (both 0.86) were on par with those illustrated in the original model development manuscript (both 0.85), but for suicide death models, we did observe a modest decline in performance (0.80–0.81 in the contemporary sample vs. 0.83–0.86 originally).

Temporal drift in calibration can be common in regression models,[20] though we observed little of that in our evaluation. Regarding discriminative ability, it is interesting that we found no meaningful change for suicide attempt models when compared with the previous publication but did for suicide death models. Suicide death coding was not impacted by transition to ICD-10-CM, and suicide death prediction models did not include any ICD-code based predictors that did not also appear in suicide attempt models (Appendix 9B–E from prior work[5]). Decline in performance for the prediction of suicide death did seem to vary across health systems (data not shown), though this variability may reflect the rarity of the outcome in our sample, especially in the smaller health systems. Suicide death outcomes are notably rarer than suicide attempts, which would have made the original model development for predicting suicide death more prone to overfitting than model development for predicting suicide attempt.

It is possible, though, that predictor distributions may have changed in a manner that differentially impacted performance of the suicide death models. The suicide death models include notably fewer predictors (43 for mental health specialty and 29 for general medical) than suicide attempt models (94 and 102, respectively), and the PHQ-9 based predictors are of somewhat greater importance in the suicide death models, accounting for approximately a third of those models' predictors. Notably, the proportion of visits in the contemporary sample that had a PHQ-9 item 9 score recorded at the visit was more than double the proportion observed in the original sample. We had expected that increased availability of PHQ-9 data might improve performance of the models and were a bit surprised it did not. Prior work, however, had found that incorporation of PHQ-9 depression questionnaire data did not lead to dramatic improvements in predictive performance compared with models that already incorporated patient demographics and EHR measures of diagnoses, prescriptions, and utilization.[21] Further, it is possible that wider use of depression questionnaires in current clinical practice[11] may now reflect a different relationship between suicide risk and presence of depression scores relative to what was observed during the time the model was originally developed.

This points to the important role that temporal changes in clinical practice patterns may play in whether future model performance continues to reflect earlier performance,[22] an issue that may be especially relevant to models predicting rare events such as suicide death. In developing prediction models, researchers are often inclined to use a development sample spanning many years of historical clinical system data to include a sufficiently large number of events. Such approaches, however, increase the likelihood of differences in patient populations, clinical practice, and informatics environments between the model development period and the period when the model is to be utilized. Our work examined temporal validity of logistic regression-based suicide risk prediction models, but such evaluations should be done regardless of prediction methods used.[23] Further, while our evaluation focused on model transportability across time, validation of transportability across other dimensions (e.g., external clinical system, patient subpopulations, etc.) may be relevant depending on a model's planned usage, thus highlighting the challenges of EHR-based model development and the critical importance of evaluation prior to potential clinical deployment.[24,25]

Limitations of our evaluation of the suicide risk prediction models in this contemporary sample are like those encountered in prior work.[5] Our definition of suicide attempt may include a small proportion of unintentional self-harm events. Perhaps more importantly, we are likely missing suicide attempts in situations where patients do not seek medical care or when health care providers do not recognize and document self-harm events. Further, our evaluation only considered outpatient mental health specialty and general medical visits, as those types of visits were used to develop the models. This work does not inform suicide risk prediction for people not receiving mental health treatment or people with no recorded mental health related diagnoses. Our evaluation of these risk prediction models was performed in the same health systems that contributed data to develop the models. Thus, our evaluation cannot speak to performance in different systems. We note, too, that while the contemporary sample used for model evaluations did not include any of the visits that had been used to train or validate those models originally, it did contain visits from some of the same people who had contributed visits to the original training data.

Conclusion

EHR-based risk prediction models developed for use in clinical systems should be periodically evaluated, so health systems can determine whether use of existing models is supported or whether new ones should be developed. In our evaluation of a published set of EHR-based suicide risk models for identifying patients at high risk for self-harm, we found little evidence of temporal deterioration in model performance. Overall, the models in a contemporary sample performed mostly as expected based on previous work but with a modest drop in performance of suicide death prediction. Models were well calibrated with solid discriminative ability for the visits of interest, suggesting that the models developed on data from an earlier time period at these health systems can be used to help inform clinical practice and patient care today.

Clinical Relevance Statement

Temporal changes in health systems and electronic health record (EHR) coding practices necessitate ongoing evaluation of EHR-based clinical risk prediction models. Our evaluation of previously developed EHR-based suicide risk models on more contemporary health system data determined that these models continue to demonstrate good calibration and discrimination. These models can be used to help identify patients at high risk for self-harm within these systems today.

Multiple Choice Questions

1. Evaluation of a previously developed EHR-based risk prediction model on data more proximal to the time of clinical implementation is:
   a. Important because statistical methods may have changed since the original model development.
   b. Important because health systems and EHR coding may have changed since the original model development.
   c. Not important because current model performance should not have changed since the original model development.
   d. Not important because the model will be useful regardless of changes since the original model development.

   Correct Answer: The correct answer is option b. The time elapsed between when a risk prediction model is developed and when it is implemented within a health system is often substantial, and temporal changes in health systems and EHR coding practices could result in altered relationships between the EHR-based predictors and outcomes of interest.

2. Which statement represents findings from our evaluation of the suicide risk prediction models?
   a. Both suicide attempt and suicide death models were poorly calibrated in the contemporary sample.
   b. Both suicide attempt and suicide death models had poor discriminative performance in the contemporary sample.
   c. Suicide attempt models had better discriminative performance than suicide death models in the contemporary sample.
   d. Suicide death models had better discriminative performance than suicide attempt models in the contemporary sample.

   Correct Answer: The correct answer is option c, as evidenced by ►Fig. 1 and ►Table 4 and the following results. Among mental health specialty and general medical visits, respectively, c-statistics were 0.862 (95% CI:

0.860–0.864) and 0.864 (95% CI: 0.860–0.869) for suicide attempt, and 0.806 (95% CI: 0.790–0.822) and 0.804 (95% CI: 0.782–0.829) for suicide death.

Protection of Human and Subjects Protections

This research study was performed in compliance with the responsible institutional review boards for each health system that approved use of de-identified records data for the study.

Funding

This work was supported by cooperative agreements U19 MH092201 and U19 MH121738 with the National Institute of Mental Health.

Conflict of Interest

In the prior 36 months, the senior author, G.E.S., has received consulting fees from UpToDate publishing. S.M.S. has been a coinvestigator on KPWHRI projects funded by Syneos Health, who is representing a consortium of pharmaceutical companies carrying out FDA-mandated studies regarding the safety of extended-release opioids. K.J.C. has served as a standing study section reviewer for NIH grant proposals and as an advisory board member for the Depression and Bipolar Support Alliance.

References

1 Simon GE, Coleman KJ, Rossom RC, et al. Risk of suicide attempt and suicide death following completion of the Patient Health Questionnaire depression module in community practice. J Clin Psychiatry 2016;77(02):221–227
2 Barak-Corren Y, Castro VM, Javitt S, et al. Predicting suicidal behavior from longitudinal electronic health records. Am J Psychiatry 2017;174(02):154–162
3 Choi SB, Lee W, Yoon JH, Won JU, Kim DW. Ten-year prediction of suicide death using Cox regression and machine learning in a nationwide retrospective cohort study in South Korea. J Affect Disord 2018;231:8–14
4 Kessler RC, Hwang I, Hoffmire CA, et al. Developing a practical suicide risk prediction model for targeting high-risk patients in the Veterans health Administration. Int J Methods Psychiatr Res 2017;26(03):e1575
5 Simon GE, Johnson E, Lawrence JM, et al. Predicting suicide attempts and suicide deaths following outpatient visits using electronic health records. Am J Psychiatry 2018;175(10):951–960
6 Walsh CG, Ribeiro JD, Franklin JC. Predicting risk of suicide attempts over time through machine learning. Clin Psychol Sci 2017;5(03):457–469
7 Agniel D, Kohane IS, Weber GM. Biases in electronic health record data due to processes within the healthcare system: retrospective observational study. BMJ 2018;361:k1479
8 Andreu-Perez J, Poon CC, Merrifield RD, Wong ST, Yang GZ. Big data for health. IEEE J Biomed Health Inform 2015;19(04):1193–1208
9 Pérez-Benito FJ, Sáez C, Conejero JA, Tortajada S, Valdivieso B, García-Gómez JM. Temporal variability analysis reveals biases in electronic health records due to hospital process reengineering interventions over seven years. PLoS One 2019;14(08):e0220369
10 Rockenschaub P, Nguyen V, Aldridge RW, Acosta D, García-Gómez JM, Sáez C. Data-driven discovery of changes in clinical code usage over time: a case-study on changes in cardiovascular disease recording in two English electronic health records databases (2001-2015). BMJ Open 2020;10(02):e034396
11 Rossom RC, Simon GE, Beck A, et al. Facilitating action for suicide prevention by learning health care systems. Psychiatr Serv 2016;67(08):830–832
12 Ross TR, Ng D, Brown JS, et al. The HMO research network virtual data warehouse: a public data model to support collaboration. EGEMS (Wash DC) 2014;2(01):1049
13 Charlson M, Szatrowski TP, Peterson J, Gold J. Validation of a combined comorbidity index. J Clin Epidemiol 1994;47(11):1245–1251
14 Tibshirani R. Regression shrinkage and selection via the lasso. J R Stat Soc Series B Stat Methodol 1996;58(01):267–288
15 Zeger SL, Liang KY. Longitudinal data analysis for discrete and continuous outcomes. Biometrics 1986;42(01):121–130
16 R Core Team. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria2019
17 Sing T, Sander O, Beerenwinkel N, Lengauer T. ROCR: visualizing classifier performance in R. Bioinformatics 2005;21(20):3940–3941
18 Lenert MC, Matheny ME, Walsh CG. Prognostic models will be victims of their own success, unless…. J Am Med Inform Assoc 2019;26(12):1645–1650
19 Toll DB, Janssen KJ, Vergouwe Y, Moons KG. Validation, updating and impact of clinical prediction rules: a review. J Clin Epidemiol 2008;61(11):1085–1094
20 Davis SE, Lasko TA, Chen G, Siew ED, Matheny ME. Calibration drift in regression and machine learning models for acute kidney injury. J Am Med Inform Assoc 2017;24(06):1052–1061
21 Simon GE, Shortreed SM, Johnson E, et al. What health records data are required for accurate prediction of suicidal behavior? J Am Med Inform Assoc 2019;26(12):1458–1465
22 Ghassemi M, Naumann T, Schulam P, Beam AL, Chen IY, Ranganath R. A review of challenges and opportunities in machine learning for health. AMIA Jt Summits Transl Sci Proc 2020;2020:191–200
23 Su C, Xu Z, Pathak J, Wang F. Deep learning in mental health outcome research: a scoping review. Transl Psychiatry 2020;10(01):116
24 Goldstein BA, Navar AM, Pencina MJ, Ioannidis JP. Opportunities and challenges in developing risk prediction models with electronic health records data: a systematic review. J Am Med Inform Assoc 2017;24(01):198–208
25 Van Calster B, Wynants L, Timmerman D, Steyerberg EW, Collins GS. Predictive analytics in health care: how can we know it works? J Am Med Inform Assoc 2019;26(12):1651–1654