Journal of the American Medical Informatics Association, 26(12), 2019, 1458–1465
doi: 10.1093/jamia/ocz136
Advance Access Publication Date: 16 September 2019
Research and Applications

Research and Applications

What health records data are required for accurate prediction of suicidal behavior?

Gregory E Simon,[1] Susan M Shortreed,[1] Eric Johnson,[1] Rebecca C Rossom,[2] Frances L Lynch,[3] Rebecca Ziebell,[1] and Robert B Penfold[1]

[1]Kaiser Permanente Washington Health Research Institute, Seattle, Washington, USA, [2]HealthPartners Institute, Minneapolis, Minnesota, USA, and [3]Center for Health Research, Kaiser Permanente Northwest, Portland, Oregon, USA

Corresponding Author: Gregory Simon, MD MPH, Kaiser Permanente Washington Health Research Institute, 1730 Minor Ave. #1600, Seattle, WA 98101, USA; simon.g@ghc.org

Received 29 January 2019; Revised 10 June 2019; Editorial Decision 9 July 2019; Accepted 19 July 2019

ABSTRACT

Objective: The study sought to evaluate how availability of different types of health records data affect the accuracy of machine learning models predicting suicidal behavior.

Materials and Methods: Records from 7 large health systems identified 19 061 056 outpatient visits to mental health specialty or general medical providers between 2009 and 2015. Machine learning models (logistic regression with penalized LASSO [least absolute shrinkage and selection operator] variable selection) were developed to predict suicide death (n = 1240) or probable suicide attempt (n = 24 133) in the following 90 days. Base models were used only historical insurance claims data and were then augmented with data regarding sociodemographic characteristics (race, ethnicity, and neighborhood characteristics), past patient-reported outcome questionnaires from electronic health records, and data (diagnoses and questionnaires) recorded during the visit.

Results: For prediction of any attempt following mental health specialty visits, a model limited to historical insurance claims data performed approximately as well (C-statistic 0.843) as a model using all available data (C-statistic 0.850). For prediction of suicide attempt following a general medical visit, addition of data recorded during the visit yielded a meaningful improvement over a model using all data up to the prior day (C-statistic 0.853 vs 0.838).

Discussion: Results may not generalize to setting with less comprehensive data or different patterns of care. Even the poorest-performing models were superior to brief self-report questionnaires or traditional clinical assessment.

Conclusions: Implementation of suicide risk prediction models in mental health specialty settings may be less technically demanding than expected. In general medical settings, however, delivery of optimal risk predictions at the point of care may require more sophisticated informatics capability.

Key words: suicide, machine learning, risk prediction, insurance claims, electronic health records, patient-reported outcomes

INTRODUCTION

Suicide mortality rates in the United States have increased by 25% since 2000, now accounting for over 45 000 deaths per year.[1] Nonfatal suicide attempts lead to almost 500 000 emergency department visits and 200 000 hospitalizations annually.[2] Half of people dying by suicide and two-thirds of people surviving suicide attempts received some mental health diagnosis or treatment during the prior year.[3,4] Outpatient mental health visits, therefore, are a potential occasion for preventive interventions.[5]

Effective secondary or selective prevention depends on accurate identification of people at risk. Unfortunately, traditional clinical detection of suicide risk is hardly better than chance.[6] While self-report measures (eg, the Patient Health Questionnaire-9 [PHQ-9] depression questionnaire) can accurately identify people at increased

risk,[7,8] those tools have only moderate sensitivity and moderate accuracy for identifying those at highest risk.

Several recent efforts have used health records data to develop models predicting suicide attempt or suicide death.[9–14] Some of these models have achieved overall classification accuracy (C-statistic) exceeding 80%, significantly improving on both traditional clinical assessments and self-report questionnaires in both sensitivity and ability to accurately predict high risk. Use of risk prediction tools to inform outreach programs is underway in the Veterans Health Administration[15] and planned in several civilian integrated health systems.

These prediction models have typically been developed in integrated health systems with access to comprehensive insurance claims and, in some cases, data from electronic health records (EHRs). Developing models using the richest possible data aims to maximize accuracy of prediction, but it may impede widespread implementation. Some healthcare systems hoping to use risk prediction models to direct or prompt preventive interventions may lack some data elements included in these prediction models. Some health systems with access to data on the full range of predictors may lack the capacity to update risk predictions using real-time data.

We report here secondary analyses from previously published research[16] examining the contributions of specific data types to accuracy of models predicting suicidal behavior following outpatient visits. We examine how accuracy of prediction might vary depending on availability of specific data sources (ie, insurance claims alone vs claims plus data available only from EHRs) and timeliness of data (ie, including or excluding data recorded during a visit for which a prediction is generated). Questions regarding the added value of real-time data are especially relevant to visit-based prevention efforts—delivering prompts or decision support to providers during outpatient visits.

MATERIALS AND METHODS

The study sample and methods for development and validation of risk prediction models are described in detail elsewhere[16] and summarized here. A public repository (www.github.com/MHResearchNetwork) includes specifications and code for defining predictor and outcome variables, a data dictionary and descriptive statistics for analytic datasets, code for variable selection and calibration steps, resulting model coefficients and confidence limits, and comparison of model performance in training and validation samples.

The 7 health systems participating in this research (HealthPartners; Henry Ford Health System; and the Colorado, Hawaii, Northwest, Southern California and Washington regions of Kaiser Permanente) provide insurance coverage and comprehensive health care (including general medical and specialty mental health care) to defined member populations enrolled through employer-sponsored insurance, individual insurance, capitated Medicaid or Medicare, and subsidized low-income programs. All systems recommend using the PHQ-9 depression questionnaire[17] at mental health visits and general medical visits for depression, but implementation was in progress during the study period and varied across health systems.

As members of the Mental Health Research Network, each health system maintains a research data warehouse following the Health Care Systems Research Network Virtual Data Warehouse model.[18] This resource combines data from insurance claims, EHRs, state mortality records, and census-derived neighborhood characteristics. Responsible institutional review boards for each health system approved use of these de-identified data for this research.

The study sample included outpatient visits between January 1, 2009 and June 30, 2015, by members 13 years of age or older, either to a specialty mental health provider or a general medical provider when a mental health diagnosis was recorded.

Potential predictors extracted from health system records for up to 5 years prior to each visit included demographic characteristics (age, sex, self-reported race, self-reported ethnicity, source of insurance, and neighborhood income and educational attainment), current and past mental health and substance use diagnoses (organized in 12 categories), past suicide attempts, other past injury or poisoning diagnoses, dispensed outpatient prescriptions for mental health medication (organized in 4 categories), past inpatient or emergency department mental health care, general medical diagnoses (by Charlson Comorbidity Index[19] categories), and recorded scores on the PHQ-9 (including response to Item 9 regarding suicidal ideation and responses to items 1-8 regarding other symptoms of depression).[17] Diagnosis, prescription, and service use predictors were represented as dichotomous indicators representing presence or absence of specific visit diagnosis groups, medication groups, and utilization types during specific time periods: on the day of the index visit, during the prior 90 days, during the prior year, and during the prior 5 years. To represent temporal patterns of prior PHQ-9 item 9 scores, 24 variables were calculated for each encounter to represent number of recorded PHQ-9 scores, maximum values, and modal values during 3 overlapping time periods (previous 90 days, previous 183 days, and previous 365 days). The final set of potential predictors for each encounter included 149 variables and 164 possible interactions (ie, interaction of prior suicide attempt diagnosis with sex).

Diagnoses of self-harm or probable suicide attempt were ascertained from injury or poisoning diagnoses recorded in EHRs and insurance claims accompanied by an International Classification of Diseases-Ninth Revision cause of injury code indicating intentional self-harm (E950-E958) or undetermined intent (E980-E989). Data supporting the sensitivity and positive predictive value of this definition are described in a previous publication.[16]

Suicide deaths were ascertained from state mortality records. Following common recommendations[20,21] all deaths with an International Classification of Diseases-Tenth Revision diagnosis of self-inflicted injury (X60-X84) or injury or poisoning with undetermined intent (Y10-Y34) were considered probable suicide deaths. Inclusion of injury and poisoning deaths with undetermined intent increases ascertainment of probable suicide deaths by 5%-10%.[7]

Prediction models were developed separately for mental health specialty and general medical visits, with a 65% random sample of each used for model training and 35% set aside for validation. Within each setting, separate models were estimated for any suicide attempt (fatal or nonfatal) and for suicide deaths. Models included multiple visits per person to accurately represent changes in risk within patients over time. For each visit, analyses considered any outcome in the following 90 days, regardless of a subsequent visit in between. In the initial variable selection step, separate models predicting risk of suicide attempt and suicide death were estimated using logistic regression with penalized LASSO (least absolute shrinkage and selection operator) variable selection.[22] The LASSO penalization factor selects important predictors by shrinking coefficients for weaker predictors toward zero, excluding predictors with estimated zero coefficients from the final sparse prediction model. To avoid overfitting models to idiosyncratic relationships in the training samples, variable selection used 10-fold cross-validation[23]

to select the optimal level of tuning or penalization, measured by the Bayesian information criterion.[24] In the second calibration step, generalized estimating equations[25] with a logistic link re-estimated coefficients in the training sample, accounting for both clustering of visits under patients and bias toward the null in LASSO coefficients. In the final validation step, logistic models derived from the above 2-step process were applied in the 35% validation sample to calculate predicted probabilities for each visit. Variable selection analyses were conducted using the GLMNET[26] and Foreach[27] packages for R statistical software, version 3.4.0 (R Foundation for Statistical Computing, Vienna, Austria). Confidence intervals for C-statistics were calculated via bootstrap with 10 000 replications. Results are reported as receiver-operating characteristic curves[28] with C-statistics[29,30] along with observed rates in prespecified strata of predicted probability.

For each of the 4 prediction scenarios (suicide attempt following mental health visit, suicide death following mental health visit, suicide attempt following general medical visit with mental health diagnosis, suicide death following general medical visit with mental health diagnosis) we developed, calibrated, and validated 4 models using increasingly detailed data. Model 1 included only age, sex, and data regarding diagnoses, prescriptions, and utilization prior to the day of the index visit—reflecting data typically available to an insurer or health plan. Model 2 added data regarding patient race, ethnicity, and neighborhood income and educational attainment—reflecting data that might be available to an insurer or health plan with linkage to available external resources. Model 3 added any available PHQ-9 data prior to the day of the index visit—reflecting data typically available in an integrated health system with access to insurance claims and EHR data. Model 4 added diagnosis and PHQ-9 data recorded on the day of the index visit—reflecting data that might inform predictions in an EHR environment capable of real-time calculation or updating of risk scores.

All technical materials (code for defining predictors and outcomes from standard data model, data dictionary, code for fitting and validating models, detailed model performance data) are available through our online repository (https://github.com/MHResearchNetwork/MHRN-Predicting-Suicide-Supplement)

RESULTS

The eligibility criteria previously identified 19 961 056 visits by 2 960 929 unique patients, including 10 275 853 mental health specialty visits and 9 685 203 general medical visits with mental health diagnoses. Characteristics of sampled visits are shown in Table 1.

Health system records and state mortality data identified 24 133 probable suicide attempts and 1240 probable suicide deaths in the study sample during the 90-day follow-up period. Among mental health specialty visits, 63 805 (0.02%) visits were followed by a probable suicide attempt, and 2383 (0.6%) visits were followed by a probable suicide death. Among general medical visits, 24 993 (0.3%) visits were followed by a probable suicide attempt, and 1301 (0.01%) visits were followed by a probable suicide death.

Receiver-operating characteristic curves in the 4 panels of Figure 1 illustrate classification performance of alternative models in each of the 4 prediction scenarios. Corresponding C-statistics or areas under the curves with 95% confidence limits are shown in Table 2. For prediction of suicide attempt following a mental health specialty visit, model 1 limited to data typically available to an insurer or health plan (age, sex, and historical utilization) had overall classification accuracy approximately equivalent to subsequent models allowed to include richer and richer data. For prediction of suicide death following a mental health specialty visit, inclusion of race, ethnicity, and neighborhood socioeconomic data (model 2) meaningfully improved prediction over model 1 limited to traditional insurance claims information. Including additional information from EHRs did not meaningfully improve prediction. For prediction of suicide attempt following a general medical visit with mental health diagnosis, a model 1 limited to data typically available to an insurer had overall classification accuracy equivalent to subsequent models allowed to consider additional sociodemographic information (model 2) or questionnaire data from EHRs (model 3). However, overall classification accuracy was significantly improved when available data were expanded to include diagnoses and questionnaire responses entered during the index visit (model 4). For prediction of suicide death following a mental health specialty visit, accuracy of classification improved slightly with each additional level of data availability. C-statistics reflect classification performance across the entire range of risk. More detailed analyses focused on calibration and classification performance at the upper end of the risk spectrum.

Calibration performance refers to accurate prediction of observed risk. Table 3 displays observed rates of suicide attempt and suicide death in different strata of predicted risk for each model in each of the 4 prediction scenarios. These percentages are analogous to positive predictive value, but for strata rather than dichotomous cutpoints. Improved performance would be indicated by agreement between predicted and observed rates within any stratum and by greater separation of rates across strata. Regarding prediction of suicide attempt following mental health specialty visits, performance was generally similar across different levels of data availability. Regarding prediction of suicide death following mental health specialty visits, performance improved meaningfully with addition of race, ethnicity, and socioeconomic data (model 1 to model 2) and was similar at subsequent steps. Regarding prediction of suicide attempt following general medical visits, performance improved moderately with addition of data from the index visit (model 3 to model 4). Regarding prediction of suicide death following general medical visits, performance improved slightly with each additional level of data availability. Positive predictive value percentages for dichotomous cutpoints are presented in Supplementary Table S3.

Classification performance refers to accurate sorting of high and low risk. Table 4 displays proportion of all events (suicide attempts or suicide deaths) occurring in different strata of predicted risk for each model in each of the 4 prediction scenarios. These percentages are analogous to sensitivity, but for strata rather than dichotomous cutpoints. Improved performance would be indicated by a larger proportion of events occurring after visits with highest scores or a smaller proportion of events occurring after visits with lowest scores. The pattern of results was similar to that described above for calibration performance. Regarding prediction of suicide attempt following mental health specialty visits, performance was generally similar across different levels of data availability. Regarding prediction of suicide death following mental health specialty visits, performance improved moderately with addition of race, ethnicity, and neighborhood socioeconomic data (model 1 to model 2). Using a percentile threshold of 90%, model 2 would identify 62% of events compared with 58% for model 1. Regarding prediction of suicide attempt following general medical visits, performance improved modestly with addition of data from the index visit (model 3 to model 4). Using a percentile threshold of 90%, model 4 would identify 61% of events compared with 59% for model 3. Regarding

Table 1. Characteristics of sampled visits to specialty mental health and general medical providers

	Mental health specialty visits (n = 10 275 853)		General medical visits (n = 9 685 203)		Visits followed by suicide death (n = 3684)		Visits followed by suicide attempt (n = 88 798)	
	n	%	n	%	n	%	n	%
Female	6 397 210	62.3	5 956 354	61.5	1337	36.3	58 251	65.6
Age								
13-17 y	1 031 932	10.0	385 948	4.0	48	1.3	16 872	19.0
18-29 y	1 721 536	16.8	1 265 442	13.1	519	14.1	22 466	25.3
30-44 y	2 684 135	26.1	2 058 564	21.3	958	26.0	21 489	24.2
45-64 y	3 775 495	36.7	3 793 229	39.2	1695	46.0	22 555	25.4
65 or older	1 062 755	10.3	2 182 023	22.5	468	12.7	5417	6.1
Diagnoses in prior 5 years								
Depressive disorder	7 602 628	74.0	5 276 619	54.5	3142	85.3	78 320	88.2
Bipolar disorder	1 371 412	13.3	561 209	5.8	866	23.5	27 083	30.5
Schizophrenia spectrum disorder	373 860	3.6	161 950	1.7	173	4.7	7903	8.9
Other psychotic disorder	518 380	5.0	275 313	2.8	409	11.1	13 142	14.8
Substance use disorder	2 955 856	28.8	1 822 479	18.8	1742	47.3	47 596	53.6
Prior self-harm diagnoses								
In prior year	241 512	2.4	69 719	0.7	626	17.0	21 480	24.2
In prior 5 y	423 227	4.1	161 088	1.7	742	20.1	27 965	31.5
PHQ-9 item 9 score recorded								
At index visit	1 012 916	9.9	480 634	5.0	523	14.2	12 432	14.0
At any visit in past year	1 781 713	17.3	723 685	7.5	833	22.6	21 578	24.3
Length of enrollment prior to visit								
1 y or more	8 767 230	85.3	8 049 100	83.1	3135	85.1	74 857	84.3
5 y or more	5 803 088	56.5	5 448 230	56.3	2089	56.7	45 731	51.5

PHQ-9: Patient Health Questionnaire-9.

prediction of suicide death following general medical visits, performance was not meaningfully different at different levels of data availability.

Raw data regarding model performance (sensitivity, specificity, negative predictive value, positive predictive value) across all possible cutpoints are provided in our online repository (https://github.com/MHResearchNetwork/MHRN-Predicting-Suicide-Supplement).

Supplementary Table S1a-d lists individual predictors selected for each of the 4 models for each of the 4 prediction scenarios. For prediction of suicide attempt following a mental health specialty visit (Supplementary Table S1a), variables selected regarding age, sex, diagnosis history, medication history, and utilization history were generally similar across the different levels of data availability. Coefficients or weights assigned to each predictor were also generally similar. Models 3 and 4 did select variables representing PHQ-9 responses at prior visits, and model 4 did select additional variables representing PHQ-9 responses at the index visit—but most of these were assigned relatively small coefficients or weights. In contrast, for prediction of suicide attempt following a general medical visit (Supplementary Table S1c), variables selected and estimated coefficients were generally similar for models 1, 2, and 3. However, model 4 selected several variables representing PHQ-9 responses at the index visit, and some of those predictors were assigned moderate weight in the final prediction model.

Supplementary Tables S2a-l show the cross-classification of risk score categories for adjacent models in each of the 4 prediction scenarios (ie, model 1 percentile categories vs model 2 percentile categories for prediction of suicide attempt following mental health specialty visit). In general, shifting among categories was greater for models predicting suicide death (Supplementary Tables S2d-f and j-l) than for models predicting suicide attempt (Supplementary Tables S2a-c and g-i). Shifting among categories was also greater when additional data elements yielded meaningful improvement in prediction (ie, Supplementary Table S2d comparing prediction of suicide death after mental health specialty visit for models including and not including data regarding race, ethnicity, and neighborhood characteristics) than when additional data had minimal effect on accuracy of prediction (ie, Supplementary Table S2a comparing prediction of suicide attempt after mental health specialty visit for models including and not including data regarding race, ethnicity, and neighborhood characteristics).

DISCUSSION

Using data from a sample of approximately 20 million visits from 7 large healthcare systems, we examined how accuracy of machine learning-derived suicide risk prediction models varied according to availability and timeliness of health records data. For prediction of suicide attempts following mental health specialty visits, we found that models limited to data typically available to an insurance carrier (ie, historical claims data) performed approximately as well as models also using up-to-the-minute diagnosis and patient-reported outcome questionnaire data from EHRs. For prediction of suicide death following mental health specialty visits, we found that models using historical claims data and basic sociodemographic information (race, ethnicity, and neighborhood characteristics) performed approximately as well as models using all available EHR data. For prediction of suicide attempts or suicide deaths following general medical visits, we found that use of EHR data, in-

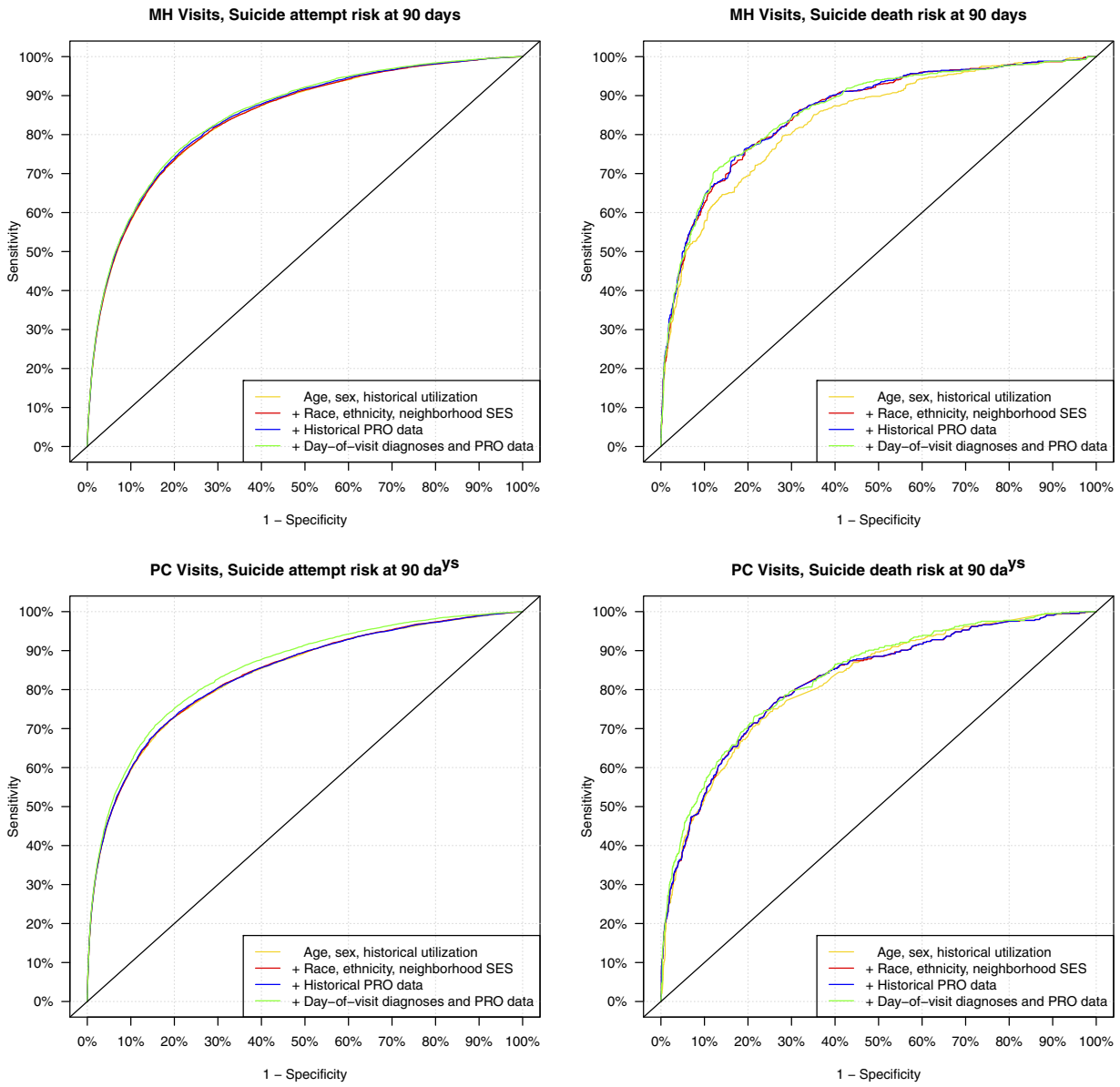

Figure 1. Receiver-operating characteristic curves illustrating model performance in validation dataset for prediction of suicide attempts and suicide deaths within 90 days of visit in 7 health systems. MH: mental health; PC: primary care; PRO: patient-reported outcome; SES: socioeconomic status.

Table 2. Overall classification accuracy for alternative models predicting suicidal behavior following outpatient visits

	Model 1		Model 2		Model 3		Model 4	
	AUC	95% CI	AUC	95% CI	AUC	95% CI	AUC	95% CI
Suicide attempt within 90 days of mental health specialty visit	0.843	(0.841-0.846)	0.843	(0.841-0.846)	0.847	(0.844-0.849)	0.850	(0.848-0.853)
Suicide death within 90 days of mental health specialty visit	0.836	(0.822-0.849)	0.859	(0.845-0.871)	0.860	(0.846-0.873)	0.861	(0.847-0.874)
Suicide attempt within 90 days of general medical visit	0.836	(0.832-0.841)	0.838	(0.833-0.843)	0.838	(0.833-0.843)	0.853	(0.848-0.857)
Suicide death within 90 days of general medical visit	0.819	(0.799-0.838)	0.821	(0.801-0.842)	0.822	(0.800-0.842)	0.833	(0.812-0.852)

Model 1 selects predictors from data regarding age, sex, and past diagnoses, prescriptions, and utilization. Model 2 selects predictors from all variables considered in model 1 and additional data regarding race, ethnicity, and neighborhood characteristics. Model 3 selects predictors from all variables considered in model 2 and additional data regarding past depression questionnaires from electronic health records. Model 4 selects predictors from all variables considered in model 3 and additional data regarding diagnoses and questionnaire responses from the index visit.

AUC: area under the receiver-operating characteristic curve; CI: confidence interval.

Table 3. Calibration performance (ie, accurate prediction of observed risk) for models using alternative data sources

Risk Score Percentile	Model 1		Model 2		Model 3		Model 4	
	Training	Validation	Training	Validation	Training	Validation	Training	Validation
Suicide attempt following mental health specialty visit								
>99.5th	12.57	12.22	12.77	12.35	12.54	12.54	13.02	12.72
99th to 99.5th	8.83	8.66	8.55	8.64	8.75	8.31	8.47	8.12
95th to 99th	4.00	4.01	4.03	4.03	4.07	4.13	4.09	4.18
90th to 95th	1.82	1.83	1.83	1.81	1.88	1.83	1.91	1.85
75th to 90th	0.85	0.85	0.85	0.86	0.85	0.85	0.86	0.86
50th to 75th	0.3	0.33	0.33	0.33	0.32	0.32	0.32	0.32
<50th	0.11	0.11	0.11	0.11	0.10	0.10	0.10	0.10
Suicide death following mental health specialty visit								
>99.5th	0.60	0.62	0.87	0.91	0.65	0.74	0.65	0.66
99th to 99.5th	0.53	0.53	0.28	0.25	0.61	0.55	0.64	0.59
95th to 99th	0.15	0.16	0.16	0.17	0.16	0.17	0.16	0.17
90th to 95th	0.07	0.06	0.08	0.08	0.07	0.08	0.07	0.09
75th to 90th	0.03	0.03	0.03	0.03	0.03	0.03	0.03	0.03
50th to 75th	0.01	0.02	0.01	0.01	0.01	0.02	0.01	0.01
<50th	<0.01	<0.01	0.01	<0.01	<0.01	<0.01	<0.01	<0.01
Suicide attempt following general medical visit with mental health diagnosis								
>99.5th	8.48	7.79	8.51	7.87	8.56	7.91	8.65	7.96
99th to 99.5th	3.88	3.98	3.75	3.93	3.80	3.98	4.16	4.19
95th to 99th	1.54	1.55	1.55	1.54	1.55	1.53	1.61	1.60
90th to 95th	0.63	0.64	0.65	0.66	0.65	0.67	0.65	0.66
75th to 90th	0.30	0.30	0.30	0.30	0.30	0.31	0.31	0.31
50th to 75th	0.13	0.13	0.13	0.13	0.13	0.13	0.13	0.13
<50th	0.05	0.05	0.05	0.05	0.05	0.05	0.04	0.04
Suicide death following general medical visit with mental health diagnosis								
>99.5th	0.28	0.19	0.48	0.42	0.48	0.42	0.54	0.44
99th to 99.5th	0.33	0.35	0.20	0.18	0.20	0.20	0.18	0.20
95th to 99th	0.09	0.08	0.08	0.08	0.08	0.07	0.09	0.08
90th to 95th	0.04	0.04	0.095	0.05	0.05	0.04	0.04	0.04
75th to 90th	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02
50th to 75th	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01
<50th	<0.01	<0.01	<0.01	<0.01	<0.01	<0.01	<0.01	<0.01

Cells display observed rates of suicide attempts and suicide deaths in different risk strata in training and validation samples.

Table 4. Classification performance (ie, accurate sorting of high and low risk) for models using alternative data sources

Risk score percentile	Model 1	Model 2	Model 3	Model 4	Model 1	Model 2	Model 3	Model 4
	Suicide attempt following mental health specialty visit (%)				Suicide death following mental health specialty visit (%)			
>99.5th	9.9	10.0	10.1	10.4	11.5	17.0	13.7	12.3
99th to 99.5th	7.0	7.0	6.6	6.4	10.0	4.7	10.3	10.8
95th to 99th	25.6	25.8	26.5	26.7	24.2	25.2	26.0	25.1
90th to 95th	14.7	14.5	14.7	14.8	11.8	15.6	14.1	16.0
75th to 90th	20.5	20.7	20.6	20.8	17.8	17.0	15.0	16.3
50th to 75th	13.4	13.4	13.2	12.9	14.5	13.5	14.1	13.6
<50th	8.9	8.7	8.3	7.9	10.2	7.1	7.0	6.0
	Suicide attempt following general medical visit (%)				Suicide death following general medical visit (%)			
>99.5th	14.9	15.1	15.1	15.3	6.3	13.9	13.7	14.4
99th to 99.5th	7.6	7.6	7.7	8.1	11.5	6.1	6.5	6.5
95th to 99th	24.1	23.8	23.6	24.7	22.0	18.4	18.4	22.2
90th to 95th	12.5	12.8	13.0	12.8	11.7	14.6	14.2	12.6
75th to 90th	17.5	17.6	17.7	18.2	22.5	22.7	22.9	19.1
50th to 75th	12.8	12.7	12.5	12.2	15.7	12.8	12.8	15.5
<50th	10.6	10.4	10.3	8.6	10.3	11.5	11.5	9.7

Cells display observed proportion of suicide attempts or suicide deaths occurring in different risk strata in training and validation samples.

cluding same-day diagnoses and questionnaire responses, yielded meaningfully more accurate predictions.

Limitations

Our measures of suicide attempt and suicide death are subject to error. As discussed previously, mortality records may not capture a significant minority of true suicide deaths. Use of recorded self-harm diagnoses to identify probable suicide attempts probably involves both false positive and false negative errors.

We only consider visits to mental health providers or visits to general medical providers with a recorded mental health diagnosis. Consequently, these models could not identify the one-third to one-half of suicide attempts and suicide deaths not preceded by any mental health diagnosis or treatment.[3,4]

Our data are all derived from integrated health systems with comprehensive records regarding receipt of general medical and mental health care, filled prescriptions for all categories of medications, and diagnoses of definite or possible self-harm in any healthcare setting. We do find that prediction models for mental health specialty care do not suffer a significant loss of performance if questionnaire data or same-day data are not available. But results may not apply to settings in which historical data are not available across all healthcare settings (primary and specialty care; outpatient, inpatient, and emergency department care) or where patients frequently receive care outside the health system.

Diagnoses assigned by real-world primary care and mental health providers are certainly subject to error and underascertainment, and more systematic or accurate diagnoses (eg, those by structured assessments or research clinicians) might yield different results.

Our findings also may not generalize to other health systems or healthcare settings with different patterns of diagnosis and treatment. For example, the associations between risk of suicidal behavior and diagnoses of mood or substance use disorder seen in most models might not be seen in settings with lower rates of recognition, diagnosis, or treatment.

Context

Consideration of more detailed or diverse predictors might lead to more accurate stratification of risk. The discrete data easily extracted from health system records do not reflect many important risk factors for suicidal behavior, including both long-term vulnerabilities (eg, adverse childhood experiences) and more proximate stressors (eg, bereavement or job loss). More systematic assessment and standardized recording of those important risk factors should be a priority for health systems hoping to accurately identify people at risk for self-harm. Analyses of full-text medical records might improve identification of standard risk factors or reveal details of providers' clinical impressions or sentiments.[12,31]

We should strongly caution against any use of these findings to infer causal relationships. Caution is especially important regarding observed associations between medication use and subsequent suicidal behavior. For example, an association between use of second-generation antipsychotic medication use and subsequent suicide attempt could reflect either a causal association (antipsychotic medication increases risk of suicide death) or, more likely, confounding (patients at higher risk of suicide death are more likely to be prescribed antipsychotic medication). Neither our data nor any other simple observational design can distinguish those 2 very different pathways.

More frequent use of the PHQ-9 might increase the gap between predictions limited to insurance claims data and predictions also including questionnaire data from EHRs, especially for general medical visits. We find that the impact of excluding PHQ-9 data was negligible for mental health specialty visits, in which such data were available for approximately 17% of encounters in our sample. In contrast, access to PHQ-9 data moderately improved prediction for general medical visits, in which such data were only available for 7.5% encounters in our sample. We suspect that the higher value of PHQ-9 data regarding general medical visits reflects sparser historical data regarding mental health diagnoses and treatment for patients seen in general medical settings. Even though PHQ-9 data were more often available regarding mental health specialty visits, those data did not add meaningfully to predictions based on richer historical data regarding diagnoses, utilization patterns, and medication use.

Even the poorest-performing models we evaluated were still superior to many of the existing risk prediction tools available. For example, prediction of suicide death following a mental health specialty visit was significantly improved by availability of data regarding race, ethnicity, and neighborhood characteristics. But a model limited to historical claims data only still yielded a C-statistic of 0.836. Using that model, observed risk of suicide death following visits with scores above the 99.5th percentile was over 50 times greater than after visits with scores below the 50th percentile. And visits with scores above the 95th percentile by that model accounted for over 45% of all suicide deaths. That performance far exceeds the accuracy of brief self-report questionnaires[6] or traditional clinical assessments.[7,8]

Implications

Implementation of suicide risk prediction models in mental health specialty clinics may be less technically demanding than expected. Our findings indicate that data typically available to an insurer or health plan (age, sex, prior diagnoses, and prior utilization patterns) can accurately predict suicide attempt or suicide death over 90 days following an outpatient visit to a specialty mental health provider. Models limited to those historical data perform well in terms of both overall classification accuracy (ie, C-statistic), sensitivity, and accurate identification of patients at highest risk. Adding questionnaire information from EHRs or diagnoses recorded at the index visit did not add significant value by any of those metrics. Consequently, health plans could consider using models developed using traditional insurance claims data to inform population-based outreach or care management programs, as already implemented by the Department of Veterans Affairs.[15,32] Healthcare delivery systems or EHR vendors could consider providing risk predictions to providers at the time of a clinical encounter, without the added complexity of recalculating risk predictions to include up-to-the-minute data.

In general medical settings, however, delivery of optimal risk predictions at the point of care may require more sophisticated EHR capability. Prediction was meaningfully improved by addition of PHQ-9 data and mental health diagnoses recorded during the index visit. Realizing that additional accuracy would require capacity to update risk predictions and deliver decision support in real time.

FUNDING

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

AUTHOR CONTRIBUTIONS

GES, SMS, EJ, RCR, FLL, RZ, and RBP were involved in study conception and design. SMS, EJ, and RZ were involved in data acquisition/analysis. GES, SMS, EJ, RCR, FLL, RZ, and RBP were involved in interpretation. GES drafted the manuscript. GES, SMS, EJ, RCR, FLL, RZ, and RBP revised the manuscript, granted final approval, and are accountable for all aspects of the work.

SUPPLEMENTARY MATERIAL

Supplementary material is available at *Journal of the American Medical Informatics Association* online.

CONFLICT OF INTEREST STATEMENT

None declared.

REFERENCES

1. Murphy SL, Xu JQ, Kochanek KD, Arias E. *Mortality in the United States, 2017*. Hyattsville, MD: National Center for Health Statistics; 2017.
2. WISQARS Nonfatal Injury Reports, 2000–2014. 2017. https://webappa.cdc.gov/sasweb/ncipc/nfirates.html Accessed April 4, 2017.
3. Ahmedani BK, Simon GE, Stewart C, *et al*. Health care contacts in the year before suicide death. *J Gen Intern Med* 2014; 29 (6): 870–7.
4. Ahmedani BK, Stewart C, Simon GE, *et al*. Racial/ethnic differences in health care visits made before suicide attempt across the United States. *Med Care* 2015; 53 (5): 430–5.
5. Patient Safety Advisory Group. Detecting and treating suicidal ideation in all settings. *Sentinel Event Alerts* 2016; 56: 1–7.
6. Franklin JC, Ribeiro JD, Fox KR, *et al*. Risk factors for suicidal thoughts and behaviors: a meta-analysis of 50 years of research. *Psychol Bull* 2017; 143 (2): 187–232.
7. 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 (2): 221–7.
8. Louzon SA, Bossarte R, McCarthy JF, Katz IR. Does suicidal ideation as measured by the PHQ-9 predict suicide among VA patients? *Psychiatr Serv* 2016; 67 (5): 517–22.
9. Kessler RC, Stein MB, Petukhova MV, *et al*. Predicting suicides after outpatient mental health visits in the Army Study to Assess Risk and Resilience in Service members (Army STARRS). *Mol Psychiatry* 2017; 22 (4): 544–51.
10. McCarthy JF, Bossarte RM, Katz IR, *et al*. Predictive modeling and concentration of the risk of suicide: implications for preventive interventions in the US Department of Veterans Affairs. *Am J Public Health* 2015; 105 (9): 1935–42.
11. Walsh CG, Ribeiro JD, Franklin JC. Predicting risk of suicide attempts over time through machine learning. *Clin Psychol Sci* 2017; 5 (3): 457–69.
12. McCoy TH, Jr., Castro VM, Roberson AM, Snapper LA, Perlis RH. Improving prediction of suicide and accidental death after discharge from general hospitals with natural language processing. *JAMA Psychiatry* 2016; 73 (10): 1064–71.
13. Kessler RC, Warner CH, Ivany C, *et al*. Predicting suicides after psychiatric hospitalization in US Army soldiers: the Army Study To Assess Risk and resilience in Service members (Army STARRS). *JAMA Psychiatry* 2015; 72 (1): 49–57.
14. Barak-Corren Y, Castro VM, Javitt S, *et al*. Predicting suicidal behavior from longitudinal electronic health records. *Am J Psychiatry* 2017; 174 (2): 154–62.
15. 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 (3): e1575.
16. 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): 951–60.
17. Kroenke K, Spitzer RL, Williams JB, Lowe B. The patient health questionnaire somatic, anxiety, and depressive symptom scales: a systematic review. *Gen Hosp Psychiatry* 2010; 32 (4): 345–59.
18. 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 (1): 2.
19. Charlson M, Szatrowski TP, Peterson J, Gold J. Validation of a combined comorbidity index. *J Clin Epidemiol* 1994; 47 (11): 1245–51.
20. Bakst SS, Braun T, Zucker I, Amitai Z, Shohat T. The accuracy of suicide statistics: are true suicide deaths misclassified? *Soc Psychiatry Psychiatr Epidemiol* 2016; 51 (1): 115–23.
21. Cox KL, Nock MK, Biggs QM, *et al.* An examination of potential misclassification of army suicides: results from the Army Study to Assess Risk and Resilience in Service members. *Suicide Life Threat Behav* 2017; 47 (3): 257–65.
22. Tibshirani R. Regression shrinkage and selection via the lasso. *J R Stat Soc Series B Stat Methodol* 1996; 58: 267–88.
23. Hastie T, Tibshirani R, Friedman J. *The Elements of Statistical Learning*. 2nd ed. New York: Springer; 2009.
24. Kass RE, Raftery AE. Bayes Factors. *J Am Stat Assoc* 1995; 90 (430): 773–95.
25. Liang JK, Zeger SL. Longitudinal data analysis using generalized linear models. *Biometrika* 1986; 73 (1): 13.
26. Friedman J, Hastie T, Tibshirani R. Regularization paths for generalized linear models via coordinate descent. *J Stat Softw* 2010; 33: 1–22.
27. Weston S. Foreach Looping Construct for R. R Package Version 1432015. https://cran.r-project.org/web/packages/foreach/index.html Accessed August 1, 2019.
28. Egan JP. *Signal Detection Theory and ROC Analysis*. New York: Springer Academic Press; 1975.
29. Hanley JA, McNeil BJ. The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology* 1982; 143 (1): 29–36.
30. Bradley AP. The use of the area under the ROC curve in the evaluation of machine learning algorithms. *Pattern Recognit* 1997; 30 (7): 1145–59.
31. Leonard Westgate C, Shiner B, Thompson P, Watts BV. Evaluation of veterans' suicide risk with the use of linguistic detection methods. *Psychiatr Serv* 2015; 66 (10): 1051–6.
32. Reger GM, McClure ML, Ruskin D, Carter SP, Reger MA. Integrating predictive modeling into mental health care: an example in suicide prevention. *Psychiatr Serv* 2019; 70 (1): 71–4.