Bringing machine learning to the point of care to inform suicide prevention

MHRN Mental Health Research Network

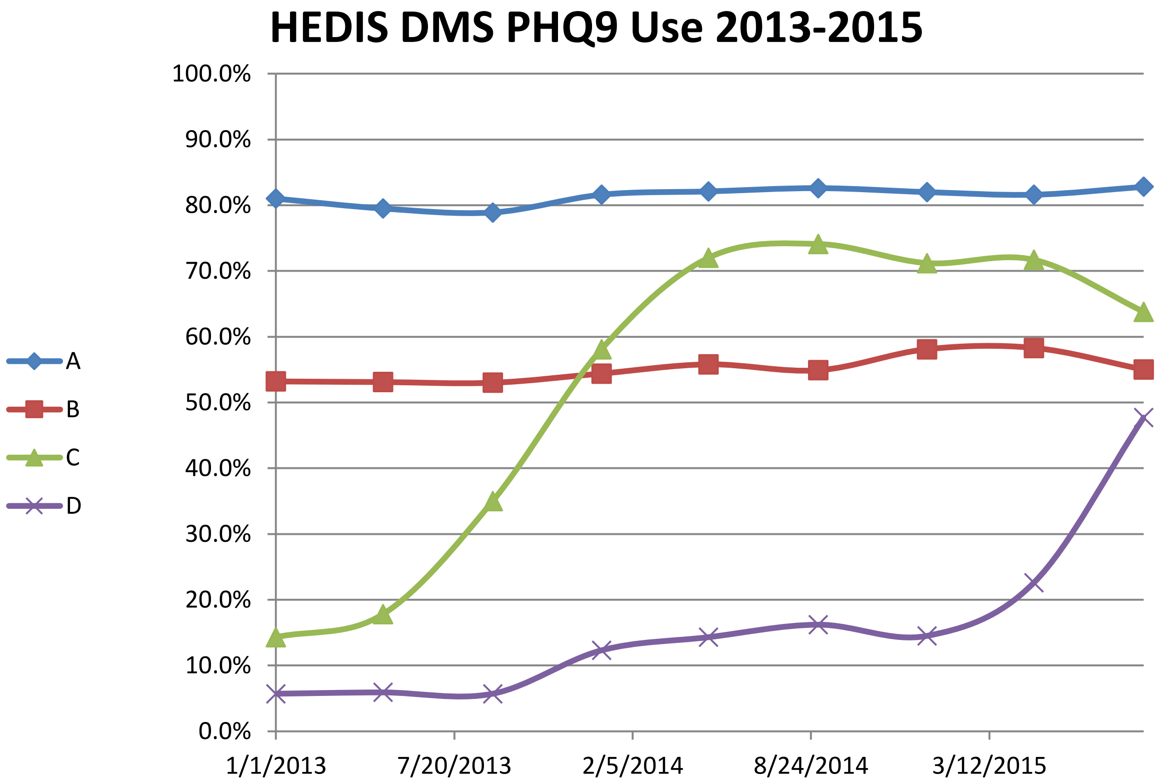
Gregory Simon and Susan Shortreed
Kaiser Permanente Washington Health Research Institute
Don Mordecai
The Permanente Medical Group

with acknowledgements to:
Eric Johnson, Rebecca Ziebell, Rob Penfold – KP Washington
Jean Lawrence – KP Southern California
Rebecca Rossom – HealthPartners
Brian Ahmedani – Henry Ford Health System
Frances Lynch – KP Northwest
Arne Beck – KP Colorado
Beth Waitzfelder – KP Hawaii

Supported by Cooperative Agreement U19 MH092201

KAISER PERMANENTE®

Measurement-based care: Uptake of PHQ9 in 4 MHRN health systems



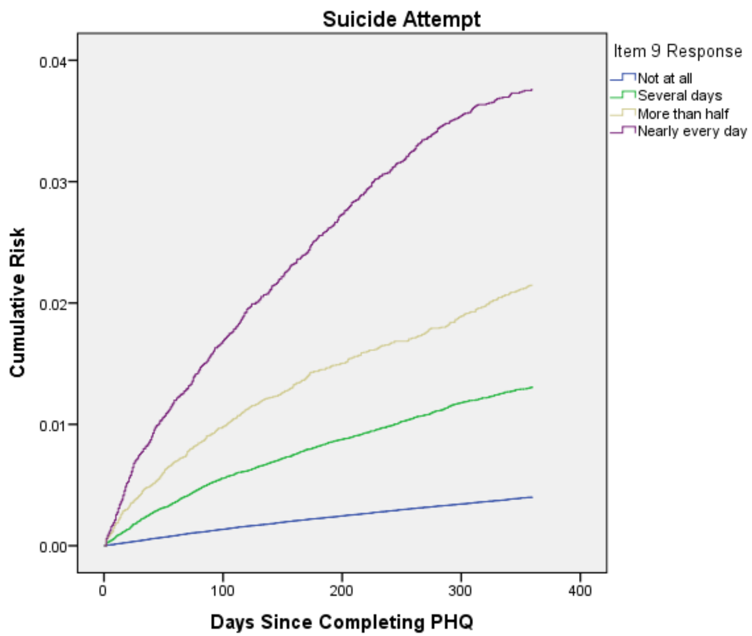
KAISER PERMANENTE®

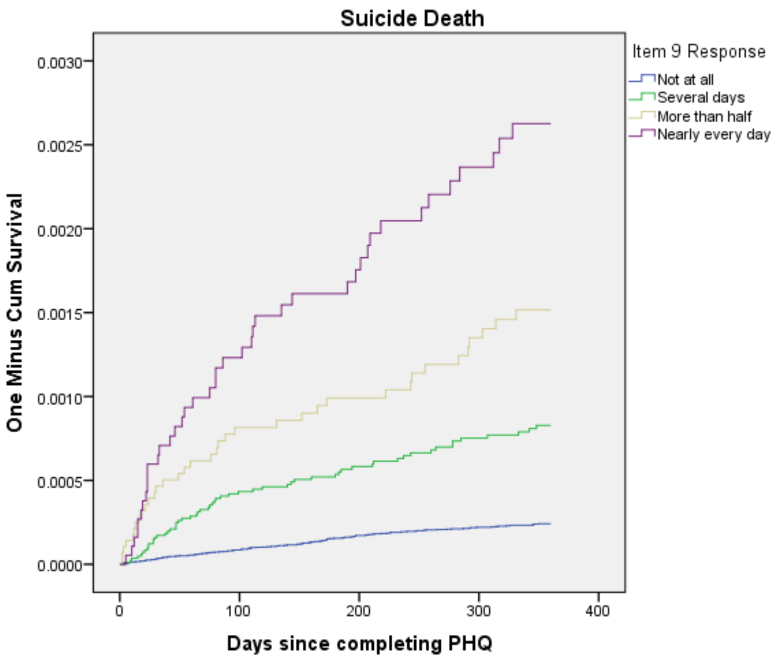
Data make new questions:

- Providers ask: What does it mean if my patient reports thoughts of death or self-harm “nearly every day”?
- Researchers answer: Nobody knows. But we could be the first to find out.

SO WE LOOKED….

Risk of suicidal behavior following completion of PHQ9





KAISER PERMANENTE®

Rapid implementation: Be careful what you wish for!

Psychiatric Services

Enter Search Term

Home Current Issue All Issues About Jobs PS in Advance Authors & Reviewers

Previous Article

Volume 64, Issue 12, December 2013, pp. 1195-1202

Next Article

Articles

Does Response on the PHQ-9 Depression Questionnaire Predict Subsequent Suicide Attempt or Suicide Death?

Gregory E. Simon, M.D., M.P.H., Carolyn M. Rutter, Ph.D., Do Peterson, M.S., Malia Oliver, B.A., Ursula Whiteside, Ph.D., Belinda Operskalski, M.P.H., and Evette J. Ludman, Ph.D.

View Author and Article Information

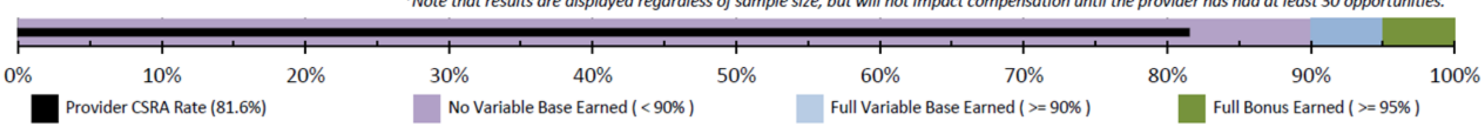
Published online: December 01, 2013 | http://dx.doi.org/10.1176/appi.ps.201200587

Quality Metric #2: BHS - 'Suicide Risk Assessment' performance through December 2015 | (0.833%)

Full variable salary is earned if a provider completes the CSRA tool for at least 90% of cases when PHQ question #9 is 2 or greater. Bonus is fully earned at 95%.

Provider	CSRA Numerator	CSRA Denominator	Provider Rate	Variable Base	Variable Base Earned	Potential Bonus Earned	Quality Metric #2 Net Impact
Simon, Gregory	31	38	81.6%	$334.25	$0.00	**$0.00**	($334.25)

**Note that results are displayed regardless of sample size, but will not impact compensation until the provider has had at least 30 opportunities.*



KAISER PERMANENTE®

Risk stratification using PHQ9 Item 9

Mental health specialty visits - Suicide attempt within 90 days

% of Visits	Item 9 Score	Actual Risk	% of Suicide Attempts
2.5%	3	2.3%	20%
3.5%	2	1.4%	19%
11%	1	0.7%	26%
83%	0	0.2%	35%

Sensitivity: 35% missed
Efficiency: Top 6% identifies 39% of events
AND – PHQ9 scores missing for significant minority of visits

KAISER PERMANENTE®

MHRN2 Suicide Risk Calculator Project

- Settings
  - 7 health systems (HealthPartners, Henry Ford, KP Colorado, KP Hawaii, KP Northwest, KP Southern California, KP Washington)
  - 8 million members enrolled
- Visit Sample
  - Age 13 or older
  - Specialty MH visit OR primary care visit with MH diagnosis
- Outcomes
  - Encounter for self-inflicted injury/poisoning in 90days
  - Death by self-inflicted injury/poisoning in 90 days

Design decisions

- Cohort design (rather than case-control)
  - Health system leaders want accurate estimation of absolute risk
  - BUT, more computationally intensive
- Sample visits (rather than people)
  - Directly inform current visit-based standard work
  - BUT, makes variance estimation more complicated
- Focus on 90-day risk (rather than longer)
  - Health system leaders ask “When can you turn off that alarm?”
  - BUT, smaller number of events reduces precision
- Use parametric (logistic) models – more later from Susan

Potential predictors

Approximately 150 indicators for each visit:

- Demographics (age, sex, race/ethnicity, neighborhood SES)
- Mental health and substance use diagnoses (current, recent, last 5 yrs)
- Mental health inpatient and emergency department utilization
- Psychiatric medication dispensings (current, recent, last 5 yrs)
- Co-occurring medical conditions (per Charlson index)
- PHQ8 and item 9 scores (current, recent, last 5 yrs)

Approximately 200 possible interactions (e.g. item 9 score WITH diagnosis of bipolar disorder)

Sample description:

- 19.6 million visits for approx. 2.9 million people
- 51% MH specialty and 48% primary care
- Race/Ethnicity: 14% Hispanic, 9% African American, 5% Asian
- Insurance source: 5% Medicaid, 20% Medicare
- Diagnoses: 1.5 million with bipolar disorder, 690K with psychotic disorders
- 1.9 million have PHQ item 9 score recorded
- 24,000 visits followed by suicide death (2108 unique events)
- 440,000 visits followed by suicide attempt (29,423 unique events)

RebeccaZiebell committed on GitHub Minor update to README ··· Latest commit b84bda9 on Jun 13

LOCAL	Initial commit of subdirectories	8 months ago
RETURN	Initial commit of subdirectories	8 months ago
README.md	Minor update to README	3 months ago
SRPM_DENOM.sas	Initial commit of SAS program	8 months ago

README.md

Suicide Risk Prediction Model (SRPM)

Denominator Programming

The Mental Health Research Network (MHRN) Suicide Risk Prediction Model (SRPM) encompasses the following major programming tasks:

1. **Identify denominator (code written in Base SAS®)**
   i. Recommended: Perform quality checks on Patient Health Questionnaire (PHQ-9) data (code written in Base SAS)
2. Create analytic data set (code written in Base SAS)
3. Implement desired model

In addition to this README, the srpm-denom repository contains the following materials that were used to perform task 1 within the MHRN.

KAISER PERMANENTE®

Teaching a computer to classify using data

- Programming requires giving the computer very specific instructions about what to do in all scenarios possible
  - Time consuming and can be very difficult
  - Especially when the set of all possible scenarios is very large
- Machine learning: let the machine learn to classify by example
  - Give the computer a set of examples already classified along with information about those examples (i.e. a training set)
    - A data set with features (variables/predictors) that describe each item
    - Identifies the correct classification of each item in the set of examples
      - Supervised learning
  - Lots of different approaches to having the computer learn from example

Machine learning to predict suicide attempts

- Goal: classify visits into those that will have and will not have a suicide attempt following the visit
  - Binary classification problem (0=no attempt, 1=attempt)
- People and health care visits have lots of “features” (predictors)
  - People: Age, sex, race/ethnicity
  - Visit:: Diagnoses, procedures, location, patient-reported outcomes (depression severity, suicidal ideation, alcohol or drug use), medications
- Give the computer some examples
  - Visits for which we know if a suicide attempt occurred in the 90 days following
  - Specify lots of features of the visits and allow machine to learn which are important for predicting which visits have a suicide attempt in the 90 days after

Selecting a machine learning method

- Used a logistic regression model for our classifier
  - Allowed the computer to select what features it used to classify
  - Created several hundred possible predictors to choose from
- Several factors impacted our selection of a parametric approach
  - Non-parametric approaches tend to be black box
    - Wanted a more transparent approach
  - Most predictors were categorical
    - Non-parametric approaches differ most in handling continuous-valued predictors
  - Anticipated parametric approach easier to implement
    - Prediction models that use simple addition and multiplication straightforward to implement within some electronic medical records systems
  - Potential protection against overfitting in a setting with rare outcomes

Tuning to prevent overfitting

- Overfitting: Good performance on the training data, but bad performance elsewhere
- A tuning parameter is often used to balance performing well on the training data and performing well in the future
  - Also called a regularization parameter
- Used Lasso to select important predictors of suicide attempt
  - Least absolute shrinkage and selection operator
  - Lasso selects predictors from a list
    - Coefficients of less powerful predictors shrunk to zero
    - Tuning parameter controls how much coefficients shrunk

Tibshirani, R. (1996). "Regression shrinkage and selection via the lasso."
J R Stat Soc Series B Stat Methodol **58**(1): 267-288.

Lasso in words, lasso in math

- Lasso selects predictors from a list
  - Coefficients of less powerful predictors shrunk to zero
    - Predictors excluded if coefficient equal to zero
  - Tuning parameter (λ) controls how much coefficient shrunk

$$\hat{\beta} = arm \min_{\beta} \underbrace{\sum_{i=1}^{n} \left(-y_i(\boldsymbol{x_i}^T \boldsymbol{\beta}) + \log(1 + e^{\boldsymbol{x}_i^T \boldsymbol{\beta}}) \right)}_{\text{Traditional MLEs for logistic regression}} + \underbrace{\lambda \sum_{j=1}^{p} |\beta_j|}_{\text{Shrinks some MLE estimates}}$$

Training, tuning, and evaluating

- Split our data (19.6 million visits) into pieces
- Training set: Used 65% of data to learn how to predict suicide attempt
  - Left 35% of the data to evaluate performance (validation set)
- Cross-validation in training set to select tuning parameter
  - 10-fold: divide training set into 10 pieces
  - Fit model with different tuning parameters on 90% of training set ten times
    - Evaluate each model's performance on the other 10% of the training set ten times
  - Select tuning parameter value that did the best in the prediction part of training
- Final model fit on all training data using selected tuning parameter
  - Use this model to predict risk of suicide attempt in the validation set
  - Evaluate performance of the predictions of this final model in the validation set

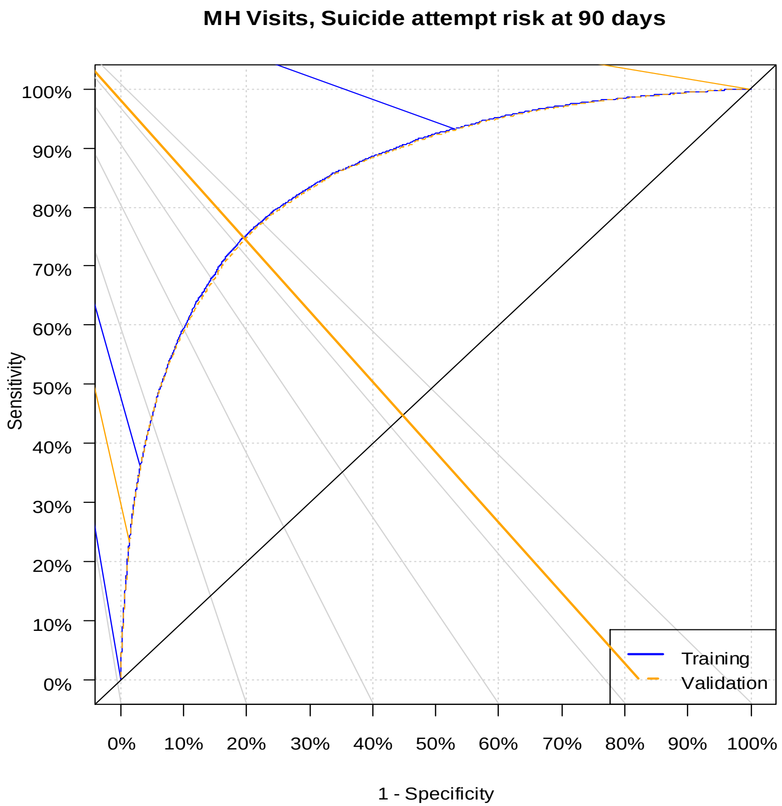
Suicidal behavior in 90 days: top 15 predictors in MH specialty care:

SUICIDE ATTEMPT FOLLOWING MH VISIT (of 94 selected)	SUICIDE DEATH FOLLOWING MH VISIT (of 62 selected)
Depression diagnosis in last 5 yrs.	Suicide attempt diagnosis in last year
Drug abuse diagnosis in last 5 yrs.	Benzodiazepine Rx. in last 3 mos
PHQ-9 Item 9 score =3 in last year	Mental health ER visit in last 3 mos
Alcohol use disorder Diag. in last 5 yrs	2^{nd} Gen. Antipsychotic Rx in last 5 years
Mental health inpatient stay in last yr.	Mental health inpatient stay in last 5 years
Benzodiazepine Rx. in last 3 mos.	Mental health inpatient stay in last 3 mos
Suicide attempt in last 3 mos.	Mental health inpatient stay in last year
Personality disorder diag. in last 5 yrs.	Alcohol use disorder Diag. in last 5 years
Eating disorder diagnosis in last 5 yrs.	Antidepressant Rx in last 3 mos
Suicide Attempt in last year	PHQ-9 Item 9 score = 3 with PHQ8 score
Mental health ER visit in last 3 mos.	PHQ-9 item 9 score = 1 with Age
Self-inflicted laceration in last year	Depression diag. in last 5 yrs. with Age
Suicide attempt in last 5 yrs.	Suicide attempt diag. in last 5 yrs. with Charlson Score
Injury/poisoning diagnosis in last 3 mos.	PHQ-9 Item 9 score = 2 with Age
Antidepressant Rx. in last 3 mos.	Anxiety disorder diag. in last 5 yrs. with Age

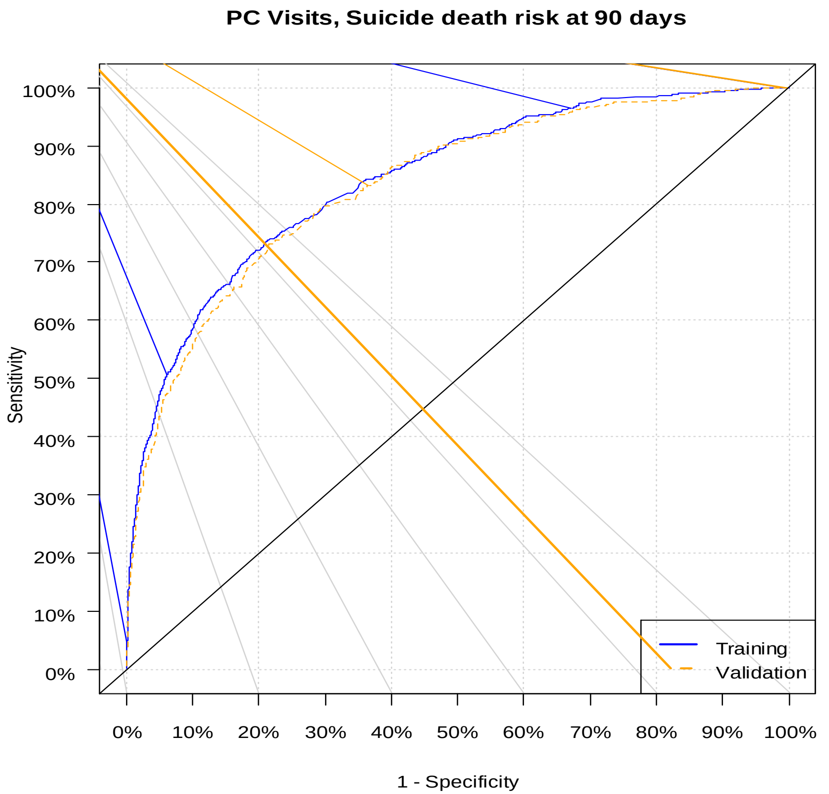
Similar predictors selected for primary care visits

KAISER PERMANENTE®

Predicting suicidal behavior in 90 days after MH visit



AUC=0.850 (0.847 - 0.853)



AUC=0.861 (0.845 - 0.877)

KAISER PERMANENTE®

AUC values for previous risk prediction models:

- Prediction of suicidal behavior:
  - Suicide death after medical hospitalization 0.74
  - Suicide death after OP visit (Army STARRS) 0.67
  - Suicide death in VA service users 0.76
  - Suicide attempt/death in health system 0.77
- Prediction of adverse medical events:
  - High ER utilization 0.71
  - Re-admission for CHF 0.62
  - In-hospital mortality after sepsis 0.76
  - Re-admission for CHF 0.78

* - no independent validation, so this is may be an over-estimate

Risk scores vs. PHQ9 Item 9 scores
Fewer events "missed" at the bottom

% of Visits	Item 9 Score	Actual Risk	% of Attempts
2.5%	3	2.3%	20%
3.5%	2	1.4%	19%
11%	1	.72%	26%
83%	0	.19%	35%

Excludes all those missing PHQ9!

% of Visits	Predicted Risk	Actual Risk	% of All Attempts
>99.5th	13.0%	12.7%	10%
99th to 99.5th	8.5%	8.1%	6%
95th to 99th	4.1%	4.2%	27%
90th to 95th	1.9%	1.8%	15%
75th to 90th	0.9%	0.9%	21%
50th to 75th	0.3%	0.3%	13%
<50th	0.1%	0.1%	8%

KAISER PERMANENTE®

Risk scores vs. PHQ9 Item 9 scores: Greater concentration of risk at the top

% of Visits	Item 9 Score	Actual Risk	% of Attempts
2.5%	3	2.3%	20%
3.5%	2	1.4%	19%
11%	1	.72%	26%
83%	0	.19%	35%

Excludes all those missing PHQ9!

Percentile of Visits	Predicted Risk	Actual Risk	% of All Attempts
>99.5th	13.0%	12.7%	10%
99th to 99.5th	8.5%	8.1%	6%
95th to 99th	4.1%	4.2%	27%
90th to 95th	1.9%	1.8%	15%
75th to 90th	0.9%	0.9%	21%
50th to 75th	0.3%	0.3%	13%
<50th	0.1%	0.1%	8%

KAISER PERMANENTE®

Using risk scores to drive standard work:

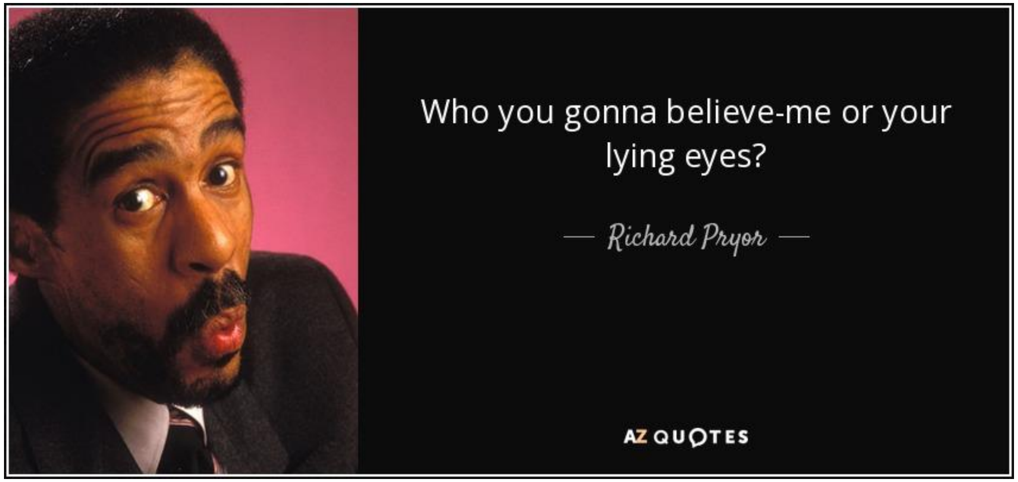
- During visits:
  - Trigger completion of CSSRS (as we do now based on PHQ9 Item 9 response)
  - Trigger creation/updating of safety plan (as we do now based on CSSRS score)
- Between visits:
  - Outreach for higher-risk patients who cancel or fail to attend scheduled visits
  - Outreach for higher-risk patients without follow-up scheduled within recommended interval

Implementation questions:

- For any threshold, risk scores are both more sensitive and more efficient than what we do now (item 9 of PHQ9).
- But…should we really ask providers to ignore item 9 responses?

Implementation questions:

- For any threshold, risk scores are both more sensitive and more efficient than what we do now (item 9 of PHQ9).
- But…should we really ask providers to ignore item 9 responses in favor of an algorithm?
- Empirical vs. Experiential knowledge: Philosophers call this “The Richard Pryor Problem”



KAISER PERMANENTE®