PATIENT TOOLS

Determining the Accuracy of Probabilistic Scoring

PATIENT TOOLS

Evidence-based Validated Measures

- Measuring latent traits (depression, anxiety, etc.)
- Severity and result regions of the latent trait
- Items with good psychometric properties (reliable, accurate)
- Original validation dataset (rater or population classification)
- Concurrent validation dataset (scored measure classification)
- Comorbidity of latent traits (depression as psych temperature)

PATIENT HEALTH QUESTIONNAIRE-9 (PHQ-9)

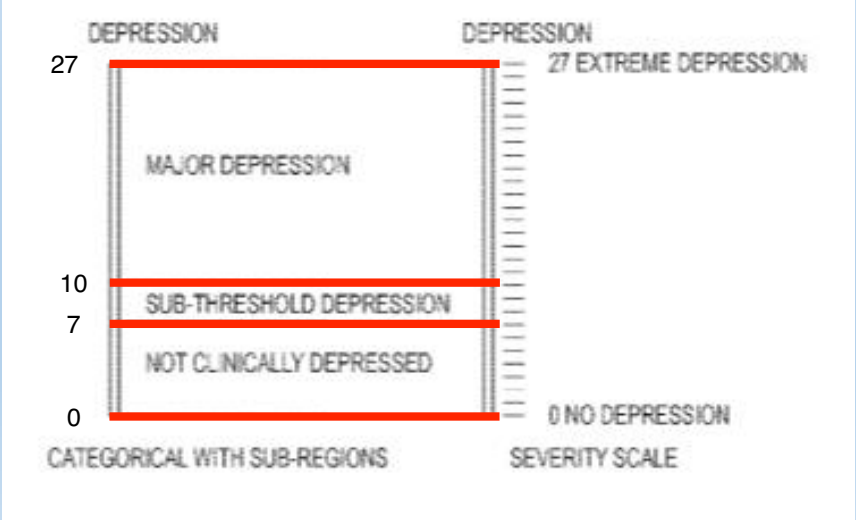
Over the last 2 weeks, how often have you been bothered by any of the following problems? *(Use "✔" to indicate your answer)*	Not at all	Several days	More than half the days	Nearly every day
1. Little interest or pleasure in doing things	0	1	2	3
2. Feeling down, depressed, or hopeless	0	1	2	3
3. Trouble falling or staying asleep, or sleeping too much	0	1	2	3
4. Feeling tired or having little energy	0	1	2	3
5. Poor appetite or overeating	0	1	2	3
6. Feeling bad about yourself — or that you are a failure or have let yourself or your family down	0	1	2	3
7. Trouble concentrating on things, such as reading the newspaper or watching television	0	1	2	3
8. Moving or speaking so slowly that other people could have noticed? Or the opposite — being so fidgety or restless that you have been moving around a lot more than usual	0	1	2	3
9. Thoughts that you would be better off dead or of hurting yourself in some way	0	1	2	3

FOR OFFICE CODING 0 + _____ + _____ + _____

=Total Score: _____

Conventional Scoring

- Linear integer value for each answer
- Compute linear severity score by summing answer values
- Select result based on range within the severity scale
- Accuracy based on sensitivity and specificity of the result



PATIENT TOOLS

Probabilistic Scoring

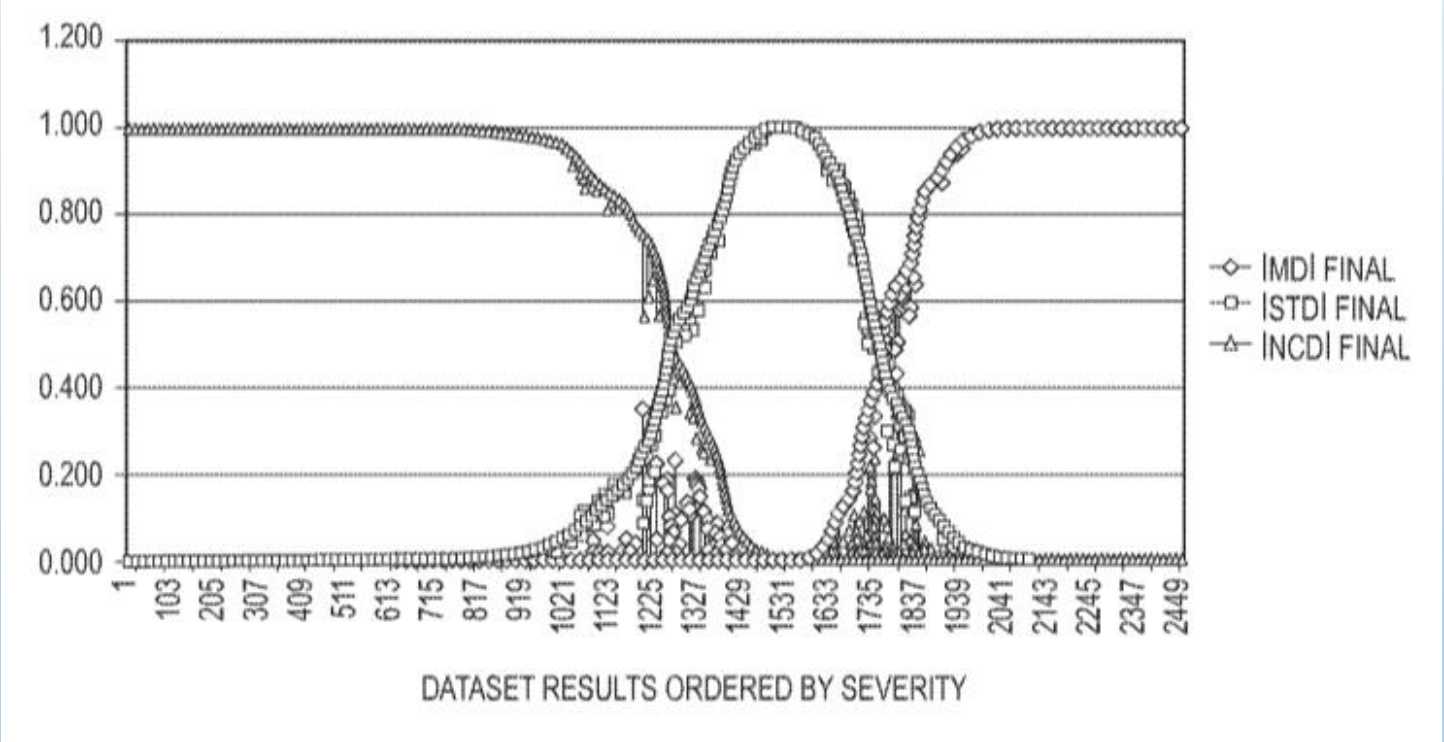
- Item Response Theory (IRT) – functional relationship, item to result
- Use Bayes Theorem – $P(C|E) = (P(E|C)/P(E)) * P(C)$
- Compute $P(E|C)/P(E)$ for each item answer in the validation dataset
- Using a lookup table, compute the result directly in real-time after each answer is given
- Dynamically administer by only asking as many questions as are needed to select a result region with adequate certainty

PATIENT TOOLS

PHQ-9 Evidence Coefficients
P(E|C)/P(E)

	Not Clincally Depressed				Sub-Threshold Depression				Major Depression			
	0	1	2	3	0	1	2	3	0	1	2	3
PHQ-1	1.535921	0.736532	0.254765	0.330620	0.494097	2.054314	1.236479	0.527800	0.301661	0.975541	2.184291	2.376285
PHQ-2	1.645223	0.701553	0.167270	0.016527	0.485251	2.332852	0.714763	0.188321	0.116167	0.909036	2.574183	3.075907
PHQ-3	1.679384	1.069410	0.451923	0.154546	0.382142	1.880743	1.397759	0.874984	0.104003	0.477565	1.768789	2.523103
PHQ-4	1.697888	1.136584	0.473656	0.161893	0.332800	1.653902	1.337789	0.792300	0.094443	0.464641	1.758481	2.548111
PHQ-5	1.547740	0.876037	0.185321	0.129121	0.569675	2.103316	1.106140	0.472926	0.246663	0.711266	2.364211	2.750739
PHQ-6	1.517303	0.569349	0.036616	0.016880	0.661431	2.531790	0.625862	0.235097	0.257552	1.047459	2.841324	3.053939
PHQ-7	1.452866	0.530565	0.116992	0.019785	0.739826	2.273619	0.525170	0.225445	0.333498	1.232574	2.747919	3.053310
PHQ-8	1.346661	0.347380	0.102221	0.039712	0.775626	2.247434	0.388266	0.150837	0.501317	1.562172	2.836033	3.052816
PHQ-9	1.185392	0.315133	0.026666		0.835730	2.162279	0.506434		0.753520	1.656964	2.913099	3.190537

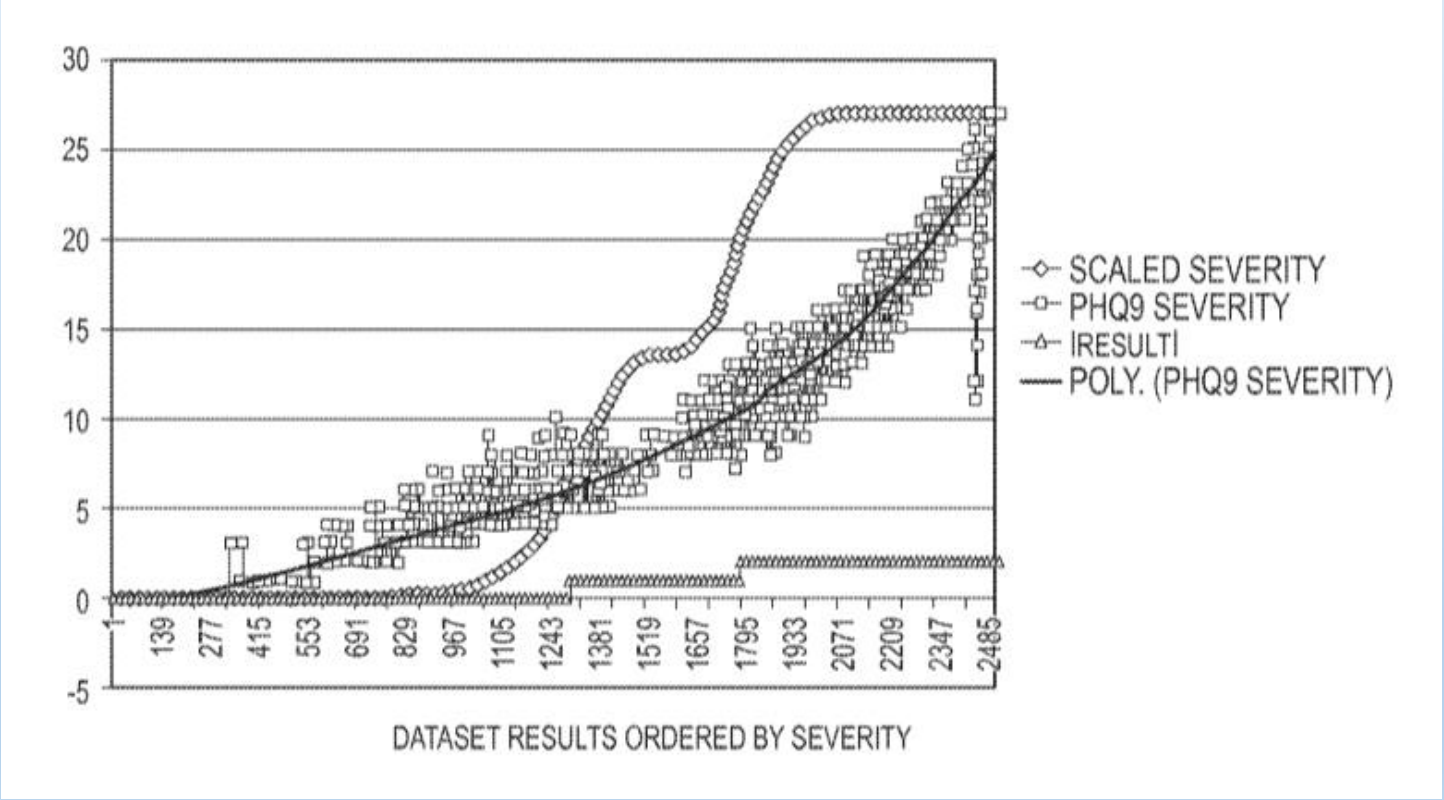
PHQ-9 Result Probabilities



(n=2495)

PATIENT TOOLS

Estimated PHQ-9 Severity from Differential Probabilities



(n=2495)

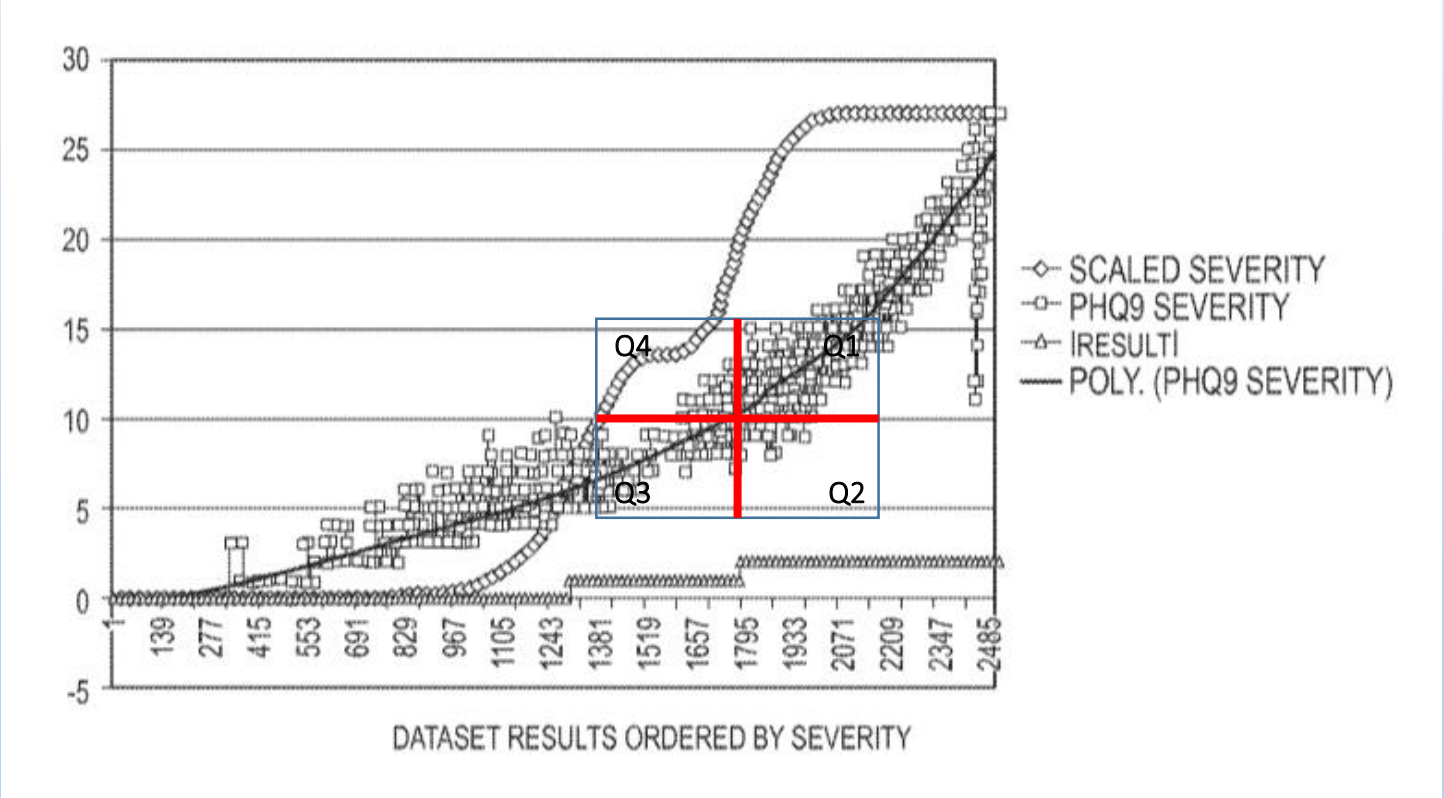
PATIENT TOOLS

Accuracy Matters

- False-positives waste time and resources for additional evaluation
- False-negatives miss issues and creates cost for the system
- Using too many question items creates screening fatigue and limits breadth

PATIENT TOOLS

PHQ-9 Conventional Scoring Error



Q1 – Sensitivity (88%)
Q2 – False-negative
Q3 – Specificity (88%)
Q4 – False-positive

(n=2495)

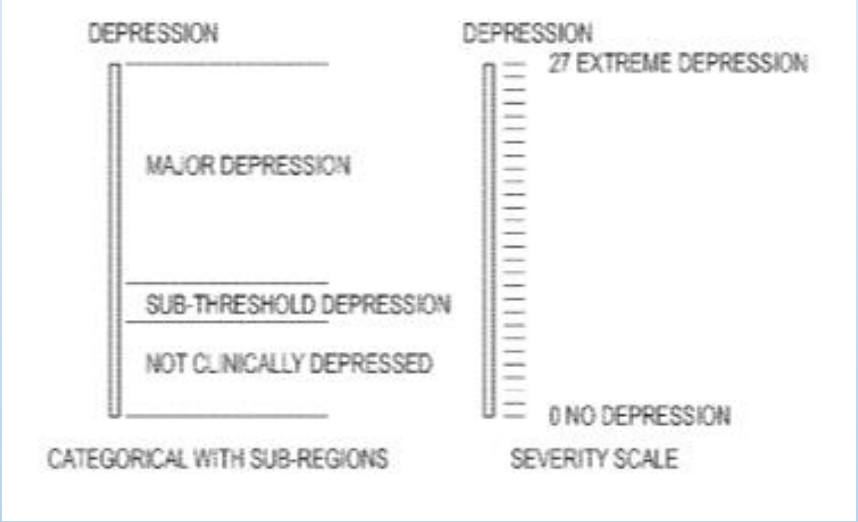
PATIENT TOOLS

The Problem At Hand

- First mathematically prove that probabilistic scoring is more accurate than conventional scoring
- Second mathematically prove that probabilistic scoring derived from a conventional scored validation dataset is essentially as accurate as using the original validation dataset and therefore still more accurate than conventional scoring

Latent Trait Scoring Error As Positional Error

- Treat knowledge as an n-dimensional space (John Ware's measuring stick)
- Conventional scoring is a linear sum of integer item values of equal unit length with a result being in an integer range
- Probabilistic scoring measures the result directly and computes severity based on differential probabilities



PATIENT TOOLS

Sources of Error

- Rater
- Person
- Using integer unit length in conventional scoring
- Treating a latent trait as a linear sum of integer value items
- Sampling error using conventional scoring versus the original rater
- ???

PATIENT TOOLS

An Engineer's Solution

- Probabilistic scoring is more accurate than conventional scoring because it bypasses error due to integer unit length and linear sums of integer item values
- The error in a conventionally scored dataset is primarily measurement sampling error due to linear, unit length scoring
- The sampling error in computing P(E) and P(E|C) is a normal distributions, as the sample size goes up, the error is driven down (this could also cover some rater and person error as well)
- The item itself only has so much predictive information that comes out in the difference in P(E|C)/P(E) coefficients

How would you solve this?

PATIENT TOOLS