Comparing Accuracy

Traditional vs Probabilistic Scoring

of PHQ9 Data

Consultation Summary Report

Client: Alan Malik, PhD - Patient Tools, inc.

Consultant: Lee Panter - University of Colorado - Denver

Contents

1 Introduction

1.1 Background

The Patient Health Questionnaire-Nine (PHQ9) is a nine-question module used for screening, monitoring and grading depressive symptions related to criteria outlined by the Diagnostic and Statistical Manual of Mental Health Disorders (DSM-IV) [1]. The PHQ9 can be administered by medical staff, or it can be self-administered either electronically or in paper format. Responses to the nine questions correspond to a numerical value and a written description of frequencies pertaining to activities, feelings, and thoughts related to depression symptoms. The PHQ9 was originally developed by Dr. Robert J. Spitzer, Dr. Janet B.W. Williams, Dr. Kurt Kroenke in 1999 with a grant from Pfizer [2].

Responses to the PHQ9 are classified into a discrete set of groups which can be ordered according to the severity of symptoms contained within the group. This analysis considers classification of PHQ9 data into three categories with the depression risk assignment values: C_1 - "Not clinically depressed", C_2 - "Sub-threshold depression", and C_3 - "Major depression".

Classification of responses is traditionally performed using the total sum of numerical scores assigned to answers provided. Traditional classification asserts that the sum of the provided answers can be used to classify an observation in conjunction with decision threshold values. This method has been shown to achieve 88% accuracy in previous investigations [3], but it assumes that each question is contributing equally to the outcome. Traditional classification also requires test takers to answer all nine questions to obtain comparable classification sum-values. These factors have led to a significant decrease in the effective implementaion of the PHQ9 in the clinical environments it was designed to function [4]. Practitioners suffer from false-positive and false-negative results that cause patient concern and healthcare system burdens, and ultimately clinical fatigue in implementation of the PHQ9.

An alternative method of classifying PHQ9 observations is Probabilistic classification. Probabilistic classification improves unpon traditional methods by using a pre-trained algorithm to iteratively re-weight probabilities for being classified within a given depression risk category until a sufficient confidence threshold is met. The probabilistic classification algorithm allows for early-stopping when taking the PHQ9 (certain answer subsets may contain enough information to probabilistically classify without needing the all nine answers provided). Probabilistic classification is also not reliant on integer-length distances and similarity measures that limit the resolution of the traditional classification criteria. The probabilistic classification algorithm has the possibilities of gaining accuracy with training sample size and relating outcomes to actionable information and treatment plans.

This analysis seeks to compare the accuracy of traditional and probabilistic classification on provided PHQ9 data. It also seeks to investigate the relationship between probabilistic classification accuracy and training sample size, and determine how this effect interacts with the accuracy of traditional classification.

1.2 Data

The PHQ9 data that is used to conduct this analysis originated from a Federally Qualified Health Research Center in Montana, United States of America [4]. The data was collected over a six month period of time, employing electronic-tablet administration of the module.

The data sample consists of 2495 observations on 286 variables and is observational with respect to measurements made on PHQ9 outcomes. In addition to PHQ9 answer variables, the data also contains demographic information (age and gender), record-keeping variables (time, date, record numbers,...etc), and the results of another psychological evaluation that is not considered in this analysis. Observations of PHQ9 variables are integer-valued (0,1,2, and 3), corresponding to the numerical association of the response provided in the module. Nine questions constitute a full answer set, and all nine questions had the same possible

answers (0,1,2, and 3) for each question. It is assumed that questions are provided, and answered sequentially, i.e. question ordering was followed. The data has been de-indentified, and contains no missing observations.

The data contains no representation of the response. In other words, the data contains real responses to PHQ9 tests, and these responses can be traditionally or probabilistically classified into one of the three depression risk categories based upon these values; however, the data does not iclude any variable that can be used to determine which method is closer to the underlying truth. The data is self-reported, responses can be misrepresented or misinterpreted by the participant.

1.3 Classification Methods

A general framework for the traditional and probabilistic classification methods is provided. Suppose that the index $i = 1, 2, \ldots, N = 2495$ represents a particular observation from the PHQ9 data, and let $q = 1,\ 2, \ldots, 9$ represent a specific question within each response sequence. Each response, to each qustion in an oberservation sequence will be denoted $A_{iq} = a_{iq}$, where the random value A_{iq} is associated with the outcome a_{iq}. Referrences to an arbitrary (fixed) observation value $I = i^*$ will utilize the shortened notation: $A_q = A_{i^*q} = k$ where $k \in \{0,\ 1,\ 2,\ 3\}$.

1.3.1 Traditional Classification

The traditional classification of observation i will be denoted:

$$\mathbb{T}_i^c \in \mathbf{T}_i^C = \left(\mathbb{T}_i^1,\ \mathbb{T}_i^2,\ \mathbb{T}_i^3\right)$$

The taditional classification of an observation is determined using the sum of the answer set provided:

$$S_i = \sum_{q=1}^{9} a_{iq}$$

Traditional depression classification outcomes are distinguished by threshold values $\alpha_1,\ \alpha_2\ \in [0,\ 27]$ with $\alpha_1 \leq \alpha_2$.

The traditional classification outcome sets can then be defined as $\mathbb{T}^c$ for $c = 1, 2, 3$:

$$\mathbb{T}^1 = \left\{ i \;\middle|\; S_i < \alpha_1 \right\} \quad \mathbb{T}^2 = \left\{ i \;\middle|\; \alpha_1 \leq S_i < \alpha_2 \right\} \quad \mathbb{T}^3 = \left\{ i \;\middle|\; \alpha_2 \leq S_i \right\}$$

where $i = 1, \ldots, 2495$ represents a single observation from the PHQ9 data.

1.3.2 Probabilistic Classification

The probabilistic classification of observation i will be denoted:

$$\mathbb{P}_i^c \in \mathbf{P}_i^C = (\mathbb{P}_i^1,\ \mathbb{P}_i^2,\ \mathbb{P}_i^3)$$

Given a set of training observations $\mathbf{\Omega}_{Train}$ of length $N_{train} = |\mathbf{\Omega}_{Train}|$, it is possible to characterize the probability of certain events within $\mathbf{\Omega}_{Train}$.

The probability that a response value of k is provided to question q is given by:

$$P\left(A_q = k\right) = \frac{1}{N_{train}} \sum_{i=1}^{N_{train}} \mathbb{I}\left(A_{iq} = k\right)$$

where

$$\mathbb{I}\left(A_{iq} = k\right) = \begin{cases} 0 & \text{if} \quad a_{iq} \neq k \\ 1 & \text{if} \quad a_{iq} = k \end{cases}$$

Similarly, the probability that an observation is assigned (via traditional classification) into

into traditional class j where $j \in \{1,\ 2,\ 3\}$ is given by:

$$P\left(\mathbb{T}^C = j\right) = \frac{1}{N_{Train}} \sum_{i=1}^{N_{train}} \mathbb{I}\left(\mathbb{T}_i^C = j\right)$$

where

$$\mathbb{I}\left(\mathbb{T}_i^C = j\right) \begin{cases} 0 & \text{if} \quad j \neq c \\ 1 & \text{if} \quad j = c \end{cases}$$

Lastly, the probability that an observation's response value is $k \in \{0, 1, 2, 3\}$ in question $Q = q$ and that same observation also has been assigned into traditional classification class $\mathbb{T}^C = j$ is given by:

$$P\left(A_q = k \bigcap \mathbb{T}^C = j\right) = \frac{1}{|\mathbb{T}_{Train}^j|} \sum_{i \in \mathbb{T}_{Train}^j} \mathbb{I}\left(A_{iq} = k\right)$$

where we are defining the set:

$$\mathbb{T}_{Train}^j = \left\{ i \in \mathbf{\Omega}_{Train} \;\middle|\; i \in \mathbb{T}^j \right\}$$

for $j = 1, 2, 3$ and the value $|\mathbb{T}_{Train}^j|$ represents the number of elements in $\mathbb{T}_{Train}^j$

These probabilities allow for the definition of a weight parameter corresponding to each 108 distinct combination of question number ($Q = q \in \{1,\ 2, \ldots, 9\}$), response value ($A_q = k \in \{0, 1, 2, 3\}$), and traditional classification: ($\mathbb{T}^C = j \in \{1,\ 2,\ 3\}$).

$$\mathbf{W}_{A_q}^{\mathbb{T}^j} = \frac{P\left(A_q = k \;\middle|\; \mathbb{T}^C = j\right)}{P\left(A_q = k\right)} \qquad \text{(EQ: 1.3.2-1)}$$

$$= \frac{P\left(A_q = k \;\cap\; \mathbb{T}^C = j\right)}{P\left(A_q = k\right) P\left(\mathbb{T}^C = j\right)}$$

$$= \frac{\frac{1}{|\mathbb{T}^j_{Train}|} \sum_{i \in \mathbb{T}^j_{Train}} \mathbb{I}\left(A_{iq} = k\right)}{\left(\frac{1}{N_{train}} \sum_{i=1}^{N_{train}} \mathbb{I}\left(A_{iq} = k\right)\right)\left(\frac{1}{N_{Train}} \sum_{i=1}^{N_{train}} \mathbb{I}\left(\mathbb{T}^C_i = j\right)\right)}$$

$$= \frac{N^2_{Train}}{|\mathbb{T}^j_{Train}|} \times \frac{\sum_{i \in \mathbb{T}^j_{Train}} \mathbb{I}\left(A_{iq} = k\right)}{\left(\sum_{i=1}^{N_{train}} \mathbb{I}\left(A_{iq} = k\right)\right)\left(\sum_{i=1}^{N_{train}} \mathbb{I}\left(\mathbb{T}^C_i = j\right)\right)}$$

Calculating the values of $\mathbf{W}_{A_q}^{\mathbb{T}^j}$ constitutes the training portion of the probabilistic classification alogrithm which now proceeds to classifying newly introduced data using an iterative approach. A pre-specified confidence threshold value $\gamma \in (0, 1)$ determines the stopping point of the algoritm which proceeds according to the framework:

Let $\mathbf{P}^C_{(q)} = \left(\mathbb{P}^1_{(q)},\ \mathbb{P}^2_{(q)},\ \mathbb{P}^3_{(q)}\right)$ represent the q^{th} iteration of the probabilistic re-weighting algorithm, for $q = 1, \ldots, 9$

$$\left(\mathbb{P}^1_{(0)},\ \mathbb{P}^2_{(0)},\ \mathbb{P}^1_{(0)}\right) = \left(\frac{1}{3},\ \frac{1}{3},\ \frac{1}{3}\right)$$

$$\left(\mathbb{P}^1_{(q+1)},\ \mathbb{P}^2_{(q+1)},\ \mathbb{P}^3_{(q+1)}\right) = \frac{\left(\mathbb{P}^1_{(q)}\mathbf{W}^{\mathbb{T}^1}_{A_q},\ \mathbb{P}^2_{(q)}\mathbf{W}^{\mathbb{T}^2}_{A_q},\ \mathbb{P}^3_{(q)}\mathbf{W}^{\mathbb{T}^3}_{A_q}\right)}{\sum_{j=1}^{3} \mathbb{P}^j_{(q)}\mathbf{W}^{\mathbb{T}^j}_{A_q}} \qquad \text{(EQ: 1.3.2-2)}$$

for $q = 1, 2, \ldots, 9$

by defining

$$\mathbb{P}^*_q = \underset{j \in \{1,2,3\}}{max} \left\{\mathbb{P}^j_q\right\} \quad \text{for} \quad q = 1, \ldots, 9$$

and

$$q^* = \underset{q \in \{1,\ldots,9\}}{min} \left\{q \;\middle|\; \mathbb{P}^*_q > \gamma\right\}$$

then (provided that q^* exists), the probabilistic scoring classification is:

$$\begin{aligned}\mathbb{P}^*_{q^*} &= \underset{j\in\{1,2,3\}}{max}\ \left\{\mathbb{P}^j_{q^*}\right\} \\ &= \underset{j\in\{1,2,3\}}{max}\ \left\{\underset{q\in\{1,\ldots,9\}}{min}\ \left\{q \;\middle|\; \mathbb{P}^*_q > \gamma\right\}\right\}\end{aligned}$$

1.4 Consultation goals and deliverables

The initially specified goals provided by Dr. Alan Malik were:

> "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"

Over the course of this analysis, the objectives shifted to reflect the lack of response variable and an inability to simulate a comprehensive response distribution. Mathematical proof was deprioritized in favor of seeking a comparison of traditional and probabilistic classification that allowed for accuracy comparisons and demonstrated that probabilistic classification accuracy improves with training sample size. The provided data lacked sufficient information to conduct a formal evaluation of accuracy within the time allotted, and simmulated response values were used to evaluate relative accuracy measures as a substitue for realistic accuracy measures.

This analysis had a stated objective of producing deliverables for Dr. Malik's commercial use. Presentable results and evidence of outcomes that can be used to demonstrate the practical advantages of probabilistic scoring. Diagrams and other visualizations that demonstrate infomational gain when probabilistc classification is implemented, with particular interest relating outcomes to answers provided in specific questions.

2 Model and Methods

2.1 Quantifying Accuracy

Model classification accuracy is calculated as a function of predicted/fitted value and response value:

$$\text{Accuracy} = f(\mathbf{y}, \hat{\mathbf{y}}) = \frac{1}{N}\sum_{i=1}^{N} I\left(y_i = \hat{y}_i\right)$$

where $\mathbf{y} = (y_1, \ldots, y_N)$ are the response values, $\hat{\mathbf{y}} = (\hat{y}_1, \ldots, \hat{y}_N)$ are the predicted/fitted values, and

$$I\left(y_i = \hat{y}_i\right) = \begin{cases} 0 & \text{if} \quad y_i \neq \hat{y}_i \\ \\ 1 & \text{if} \quad y_i = \hat{y}_i \end{cases}$$

When response values are not available, accuracy can be calculated as a function of fitted values ($\hat{\mathbf{y}}$) and a *simulated* response ($\tilde{\mathbf{y}} = (\tilde{y}_1, \ldots, \tilde{y}_N)$), where it is assumed that $\tilde{\mathbf{y}} \approx \mathbf{y}$ but $\tilde{\mathbf{y}} \neq \mathbf{y}$

$$\tilde{\text{Accuracy}} = f(\tilde{\mathbf{y}}, \hat{\mathbf{y}}) = \frac{1}{N}\sum_{i=1}^{N} I\left(\tilde{y}_i = \hat{y}_i\right) \qquad \text{(EQ: 2.1-1)}$$

Simulated accuracy calculation can be performed for traditional and probabilistic classifications:

$$\tilde{Accuracy}_{\mathbb{T}} \leftarrow\rightarrow \tilde{Accuracy}_{\mathbb{P}}$$

$$f\left(\tilde{\mathbf{y}}, \hat{\mathbf{y}}_{\mathbb{T}}\right) \leftarrow\rightarrow f\left(\tilde{\mathbf{y}}, \hat{\mathbf{y}}_{\mathbb{P}}\right) \qquad \text{(EQ: 2.1-2)}$$

This analysis will focus on estimating various values of $\tilde{Accuracy}$ using multiple methods of generating values of $\tilde{\mathbf{y}}$

2.2 Cross Validation Processing:

Estimating $\tilde{Accuracy}$ using Cross Validation (CV) allows for the estimation of both probabilistic and traditional classifications simultaneously, and the ability to establish the relationship between training sample size and accuracy. Given a fixed value of $K \in \{1, 2, \ldots, N\}$ a partition of the original data is formed from K equal subsets, each of which has N_K observations sampled randomly without replacement from the original data, where:

$$N_K = \left\lfloor \frac{N}{K} \right\rfloor$$

There are K different ways in which $K-1$ data sets can be chosen from K sets when order does not matter:

$$\begin{aligned} \binom{k}{k-1} &= \frac{k!}{(k-1)!(k-(k-1))!} \\ &= \frac{k!}{(k-1)!} \\ &= k \end{aligned}$$

for each of these K combinations a train-test combination is created by combining $K-1$ selected sets to form the training set, and the remaining set to form the test set. Let the j^{th} train-test pairing in the partion of the full data into K sets be represented by:

$$(TR, TE)_j^K = \left(TR_j^K, TE_j^K\right)$$

for $j = 1, \ldots, K$.

Using the definition of N_K, and the fact that each training set is a union of $(K-1)$ sets of length N_K, it is possible to establish a relationship between training set length, i.e. $|TR_j^k|$ the number of observations in the j^{th} training set and the value of K from which the original

train-test partitions were formed:

$$|TR_j^k| = (K-1)N_K = (K-1)\left\lfloor \frac{N}{K} \right\rfloor$$

a similar relationship may be defined for the length of the test set:

$$|TE_j^k| = N_K = \left\lfloor \frac{N}{K} \right\rfloor$$

note that:

$$\left\lfloor \frac{N}{K} \right\rfloor \leq \frac{N}{K}$$

implies

$$\begin{aligned} |TR_j^K| + |TE_j^K| &= (k-1)\left\lfloor \frac{N}{K} \right\rfloor + \left\lfloor \frac{N}{K} \right\rfloor \\ &= K\left\lfloor \frac{N}{K} \right\rfloor \\ &\leq K\frac{N}{K} = N \end{aligned}$$

meaning that observations will be left out of the sampling process in the selection of subsets. This is due to the definition of the subsample sizes chosen using the floor function which is not injective on real valued domains.

It is a stated objective of this analysis to establish a relationship between $|TR_j^k|$ and $\tilde{accuracy}$, idealistically represented as:

$$\tilde{Accuracy} = \tilde{Accuracy}\left(|TE_j^k|\right) \qquad \text{(EQ-XX)}$$

If it can be shown that:

$$\hat{\mathbf{y}} = \hat{\mathbf{y}}\left(|TR_j^K|\right)$$

then

$$\begin{aligned} \tilde{Accuracy} &= f\left(\tilde{\mathbf{y}}, \hat{\mathbf{y}}\right) \\ &= f\left(\tilde{\mathbf{y}}, \hat{\mathbf{y}}(|TR_j^K|)\right) \\ &\Rightarrow \tilde{Accuracy} = \tilde{Accuracy}\left(|TE_j^k|\right) \end{aligned}$$

This formula establishes a connection between the value of K and $\tilde{accuracy}$. Demonstrating that by calculating the value of $\tilde{Accuracy}$ at a variety of K values, it is possible to establish the relationship between $\tilde{Accuracy}$ and $|TR_j^k|$ (i.e. Training Sample Length).

(EQ: 2.1-1) can be used to write an expression for the classification accuracy of each train-test pair within a specific value of K:

$$\tilde{Accuracy}\left(TR, TE\right)_j^k = \frac{1}{|TE_j^k|} \sum_{i=1}^{|TE_j^k|} I\left(\tilde{y}_i = \hat{y}_i\right)$$

the average of these values is used to represent the accuracy at a particular K-value

$$\begin{aligned} \tilde{accuracy}\left(TR, TE\right)^k &= \tilde{accuracy}_m\left(TR, TE\right)_\bullet^k \\ &= \frac{1}{K} \sum_{j=1}^{K} \left\{ \tilde{accuracy}\left(TR, TE\right)_j^k \right\} \\ &= \frac{1}{K} \sum_{j=1}^{K} \left\{ \frac{1}{|TE_j^k|} \sum_{i=1}^{|TE_j^k|} I\left(\tilde{y}_i = \hat{y}_i\right) \right\} \end{aligned}$$

and the process can be applied to traditional and probabilistic classification methods:

$$\tilde{accuracy}_{\mathbb{T}}\left(TR,TE\right)^{k} \leftarrow \rightarrow \tilde{accuracy}_{\mathbb{P}}\left(TR,TE\right)^{k}$$

$$\frac{1}{K}\sum_{j=1}^{K}\left\{\tilde{accuracy}_{\mathbb{T}}\left(TR,TE\right)_{j}^{k}\right\} \leftarrow \rightarrow \frac{1}{K}\sum_{j=1}^{K}\left\{\tilde{accuracy}_{\mathbb{P}}\left(TR,TE\right)_{j}^{k}\right\}$$

$$\frac{1}{K}\sum_{j=1}^{K}\left\{\frac{1}{|TE_{j}^{k}|}\sum_{i=1}^{|TE_{j}^{k}|}I\left(\tilde{y}_{i}=\hat{y}_{i}^{\mathbb{T}}\right)\right\} \leftarrow \rightarrow \frac{1}{K}\sum_{j=1}^{K}\left\{\frac{1}{|TE_{j}^{k}|}\sum_{i=1}^{|TE_{j}^{k}|}I\left(\tilde{y}_{i}=\hat{y}_{i}^{\mathbb{P}}\right)\right\}$$

where $\hat{y}_{i}^{\mathbb{P}}$ is a probabilistic classification generated outcome, and $\hat{y}_{i}^{\mathbb{T}}$ is a traditional classification generated outcome. It will also be applied to a multitude of different simulated validation outcomes

$$\begin{aligned}\tilde{accuracy}_{\mathbb{T}or\mathbb{P}}^{h}\left(TR,TE\right)^{k} &= \frac{1}{K}\sum_{j=1}^{K}\left\{\tilde{accuracy}_{\mathbb{T}or\mathbb{P}}^{h}\left(TR,TE\right)_{j}^{k}\right\}\\ &= \frac{1}{K}\sum_{j=1}^{K}\left\{\frac{1}{|TE_{j}^{k}|}\sum_{i=1}^{|TE_{j}^{k}|}I\left(\tilde{y}_{i}^{h}=\hat{y}_{i}^{\mathbb{T}or\mathbb{P}}\right)\right\}\end{aligned}$$

where the values of h will be outlined in **Section XX**. Since notation is obviously cumbersome, indices displayed will be limited only to those relevant.

The possible training set observation counts using the CV method outlined above on the data provided range between 1247 and 2492 observations. The figure also shows that for K-values greater than 1248, test set sizes are confined to one. Training set sizes were selected for analysis by sampling 50 different values of K from k=5 to k=2490. It was assumed that subsets taken from the data which were sampled randomly without replacement are at least partially representative of the population as a whole.

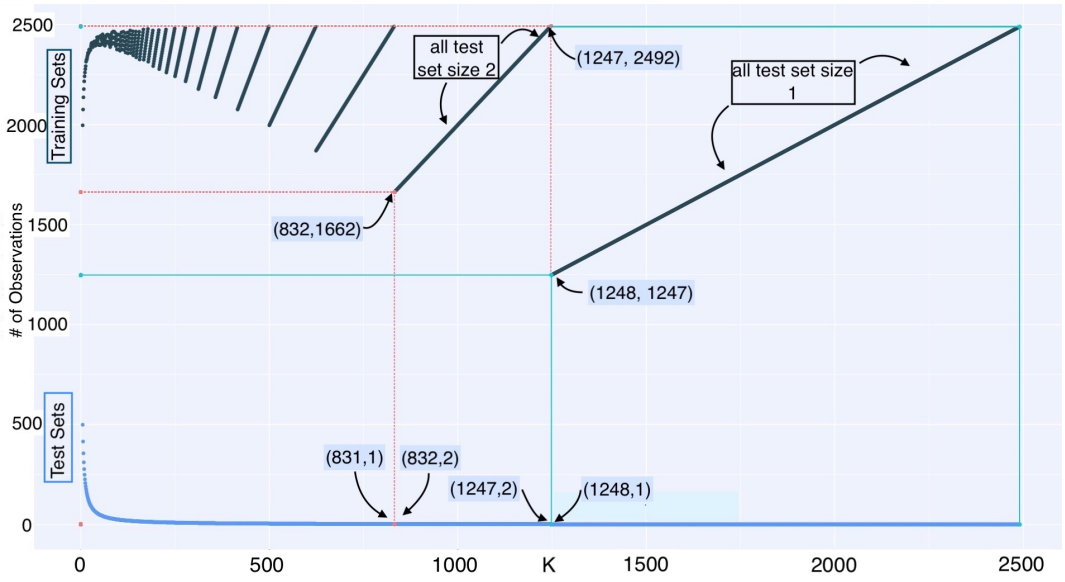

Figure 1: Training and Test set observation counts vs K

2.3 Optimal Convergence Criterion

The probabilistic classification process requires the specification of a confidence level $\gamma \in (0, 1)$ that determines the threshold at which the algorithm terminates. In this analysis, a confidence level of $\gamma = 0.75$ was chosen. This value optimizes probabilistic classification accuracy of the first simulated validation outcome (see below) when the probabilistic weights are trained on the full data set.

2.4 Simulated Validation Outcomes

2.4.1 Inferential Simulations

The first simulated outcome is generated using a series of assumptions that collectively determine a method for creating a response distribution uniquely functional on the PHQ9 data inputs and radom sampling.

The three theoretical assumptions made to generate the outcome are:

1. Depression classifications are a hierarchy, so outcomes in group 1 are generally lower than those in group 2, and outcomes in group 2 are generally lower than those in group 3
2. Within each class there is a spectrum of conditions, and transitions between classes exhibit continuous properties. This means that outcomes high in group 2 are closer in magnitude to a lower group 3 outcome compared to average or lower group 2 outcomes.
3. The 12 percent misclassification specified in **Section 1.1** is attributable to the traditional classification algorithm mis-classifying observations into a proximal class. This means that the traditional classification algorithm misclassifies according to: group 1 for group 2, group 3 for group 2, and group 2 for either group 1 or 2; however, it never mistakes group 1 for group 3 or group 3 for group 1.

The deterministic traditional classification algorithm outcome is altered using the three assumptions by implementing quantity variations related to:

- $\mathbb{S}_T = \left\{ i \;\middle|\; S_i = T \right\}$ for $i = 1, \ldots, N$ and $T = 0, 1, \ldots, 27$ where $S_i = \sum_{q=1}^{9} a_q$, and the number of observations with the same traditional sum-score (S_i) is the value: $|\mathbb{S}_T|$. Traditional sum-scores with higher observation counts will have more variation induced.
- Distance from a threshold value. Observations which are closer to a threshold value are more likely to be mis-classified, if $\alpha_1 \leq \alpha_2 \in \{0, 1, \ldots, 27\}$ are threshold values then thhe distance to the nearest threshold value of an observation i that is classified into traditional sum value S_i is:

$$D\left(\alpha_k, S_i\right) = \begin{cases} \underset{k=1,2}{min} \left\{ \left| \alpha_k - S_i \right| \right\} & \text{if} \quad \alpha_k \neq S_i \\ \frac{1}{2} & \text{if} \quad \alpha_k = S_i \end{cases}$$

The final probability of a traditional classification being altered (different from its traditional classifications) in its representation in the simulated outcome data set is

then:

$$P\left(\tilde{\mathbb{T}}_i \neq \tilde{\mathbb{T}}_i\right) \propto \frac{|\mathbb{S}_T|}{D\left(\alpha_k, S_T\right)}$$

where proportionality constants are chosen so that the total difference between Traditional Scoring classification and simulated outcome is 12%

Specific values of $P\left(\tilde{\mathbb{T}}_i \neq \tilde{\mathbb{T}}_i\right)$ that were used for this analysis are displayed in **Table XX**

$\mathbb{S}_T$	Orig Class	Sim Class	D_{α_1}	D_{α_2}	$\lvert\mathbb{S}_T^{PR}\rvert$	# Flip	Prop Const	% Total Data
5	1	2	2	5	149	42	0.282	0.017
6	1	2	1	4	137	63	0.460	0.025
7	2	1	0	3	108	42	0.389	0.017
8	2	1	1	2	115	19	0.165	0.008
8	2	3	1	2	115	21	0.183	0.008
9	2	1	2	1	134	16	0.119	0.006
9	2	3	2	1	134	21	0.157	0.008
10	3	2	3	0	72	21	0.292	0.008
11	3	2	4	1	73	21	0.288	0.008
12	3	2	5	2	77	21	0.273	0.008
							SUM →	0.113

Figure XX:

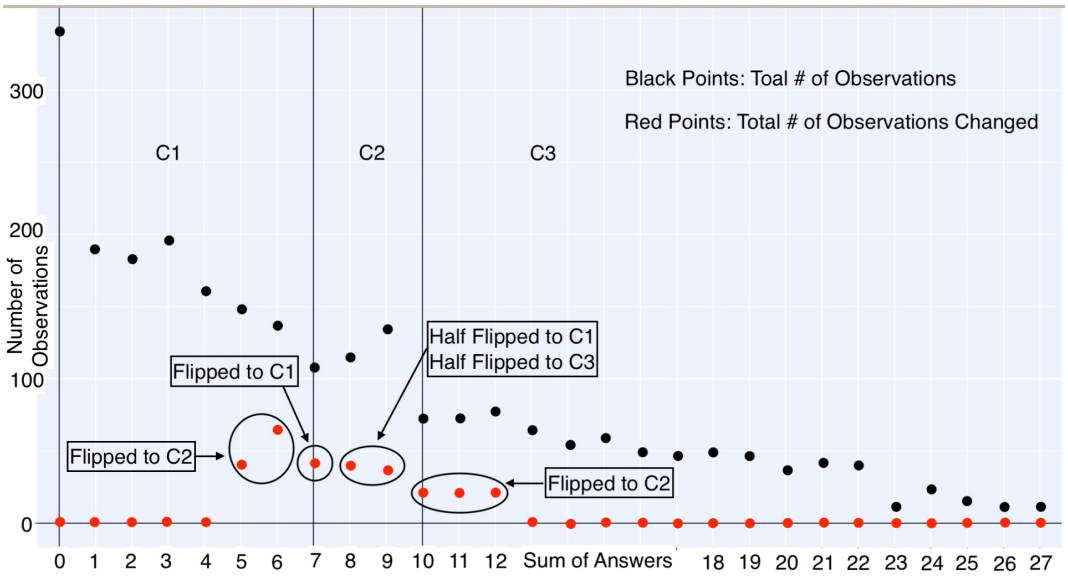

Figure XX: depicts observational distributions as classified by traditional methods, and how the first simulated validation outcome induced change

Since the simulated validation outcomes generated using this method depended only on information relevant to the observation from which the outcome was being simulated, the simulated data values are paired with their corresponding observations prior to evaluating the CV algorithm. Once paired, 50 accuracy comparisons of probabilistic and traditional classification were obtained on 50 different values of K using the CV algorithm previously outlined.

This simulated outcome is implemented in the process to obtain the optimal convergence confidence level in **Section XX**

2.4.2 Probabilistic Outcome Simulations

The probabilistic classification algorithm is appealing because it offers multiple possible advantages over traditional classification. Although not fully confirmed, intuition indicates that the accuracy of the probabilistic classification should increase with sample size. Moreover, the probabilistic classification algorithm considers the information each question response

provides independently, considering responses as sequences of information instead of an agglomeration.

Together, these aspects make probabilistic classification not only appealing for classification, but also simulation. The probabilistic algorithm incorperates information from a sample to estimate an outcome. Similar to the inferences conducted in **Section 2.4.1**, the probabilistic classification algorithm estimates can be used as a simulated responses to conduct an analysis of accuracy.

The process involves altering the CV algorithm to allow for estimation of the probabilistic and traditional classification outcomes and a distinct probabilistically simulated response, simmultanneously.

As before, the original data is separated into K-CV sets, and combined into train-test pairings

$$(TR, TE)_j^K$$

where $j = 1, \ldots, K$. For each training set, calculate the corresponding set of probabilistic training weights

$$TR_j^K \ \rightarrow \ \text{(EQ: 1.3.2-1)} \ \rightarrow \ W_{A_q}^{\mathbb{T}_j}$$

K train-test pairs means there will be K distinct estimates of the training weights. The average training weight (for each weight value)

$$W_{A_q}^{\mathbb{T}_K} = W_{A_q}^{\mathbb{T}^\bullet} = \frac{1}{K} \sum_{j=1}^{K} W_{A_q}^{\mathbb{T}^j}$$

is used for the classification of new response values:

$$\left(W_{A_q}^{\mathbb{T}_K}, TE_j^K\right) \ \rightarrow \ \text{(EQ: 1.3.2-2)} \ \rightarrow \ \tilde{y}_j^K$$

The simulated outcomes generated using this method depend on information obtained from

a training set used to generate the probabilistic weights which were used in its classifications. Each simulated outcome will therefore be paired with data that was simulated using the same set of information inside of the CV algorithm.

$$\begin{aligned} acc\tilde{u}racy_{\mathbb{T}or\mathbb{P}}\left(TR,TE\right)^{k} &= \frac{1}{K}\sum_{j=1}^{K}\left\{acc\tilde{u}racy_{\mathbb{T}or\mathbb{P}}\left(TR,TE\right)_{j}^{k}\right\} \\ &= \frac{1}{K}\sum_{j=1}^{K}\left\{\frac{1}{|TE_{j}^{k}|}\sum_{i=1}^{|TE_{j}^{k}|}\left\{I\left(\tilde{y}_{ji}^{K}=\hat{y}_{ji}^{K(\mathbb{T}or\mathbb{P})}\right)\right\}\right\} \end{aligned}$$

This method makes the assumption that the estimates of average weights are stable enough to produce reasonably consistent estimates **Figure XX:**

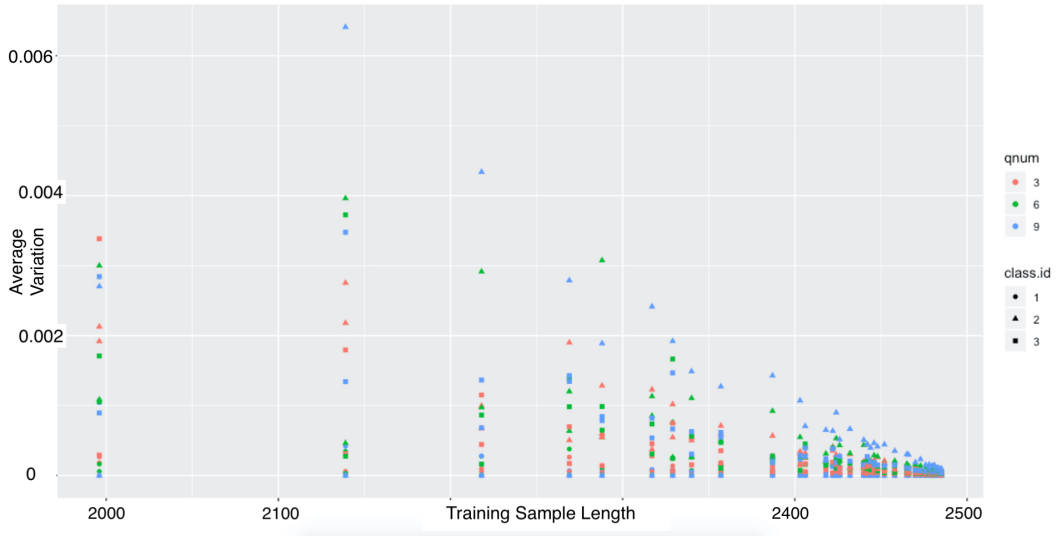

Figure XX: Variance of average weight values for several question number/class ID combinations plotted against training data count length **Figure XX:**

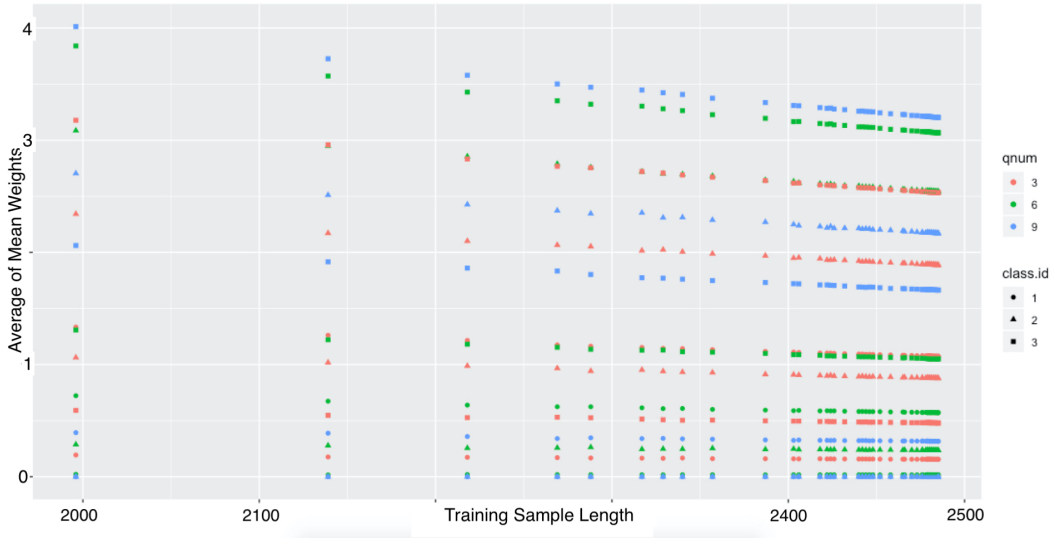

Figure XX: Mean value of the average weight estimates for several question number/class ID combinations plotted against training data count length

The stability of the average estimate value, as training sample increases, in combination with a dramatic decrease in variance as training sample size increases are indications that the average weight values are capable of producing stable outcome estimates

2.4.3 Unsupervised Learning Classifications

Unsupervised learning methods can be used to discover patterns in data without the need to specify a response variable, or include additional assumptions. The methods employed in this analysis attempt to derive trends in the PHQ9 data and infer an outcome by deriving similarities in the observations using the predictor variables, and associating similar observations with an unknown outcome variable. The ranked-outcomes $C_1,\ C_2,\ C_3$ in the PHQ9 data allow for a hierarchy to be imposed upon the unknown variables based upon the characteristics of the observations contained within each outcome. This hierarchy can then be input into the CV algorithm as a simulated outcome to obtain estimates of traditional and probabilistic classification accuracies.

2.4.3.1 Kmeans Clustering

The proper number of clusters in the Kmeans algorithm can be determined with the use of an elbow plot. The first instance of diminishing returns is two or three clusters. Since the PHQ9 data needs to be grouped into three clusters this coincides with intuition:

Figure XX:

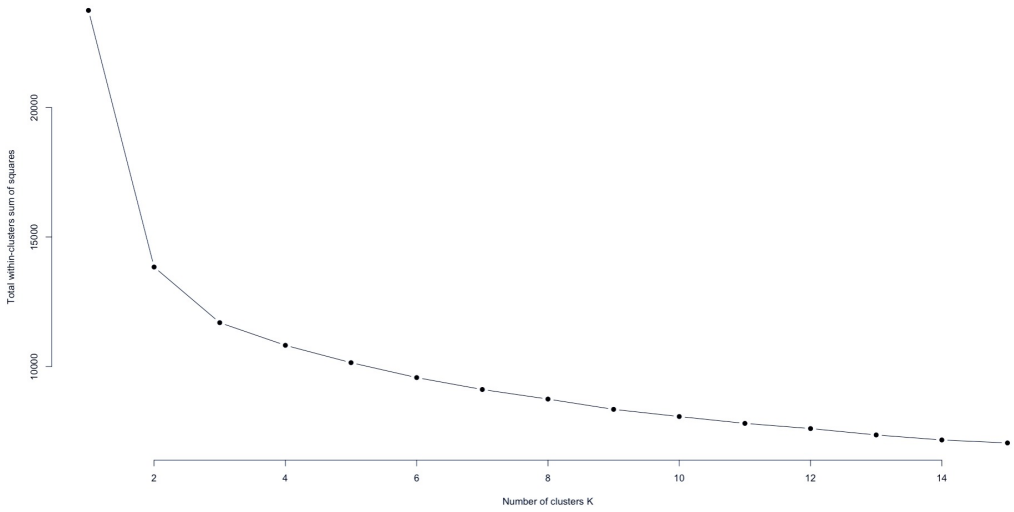

Figure XX: A plot depicting the total within sum of squares vs the number of clusters chosen for the algorithm

The Kmeans algorithm minimizes the sum of squares distance from the assigned cluster centers. A value of 2000 random intitializations with a maximum iteration per initialization of 1000 steps, and a random seed of "123" is utilized.

Cluster assignments are formed using the median traditional sum value within each cluster:

Table XX:

Cluster Label	Cluster Population	Median Traditional Sum Value
C1	1254	2
C2	776	9
C3	465	19

Table XX: Kmeans cluster assignments based upon median traditional sum value

2.4.3.2 Hierarchical Clustering

Two hierarchical clustering variables are created. The first distinguishes observations based upon Euclidean distance. A circular dendogram, with corresponding classifications is displayed in **Figure XXa** [5]. The second hierarchical clustering technique distinguishes observations based upon Manhattan distance. A circular dendogram, with corresponding classifications is displayed in **Figure XXb** [5]

Figure XXa/b:

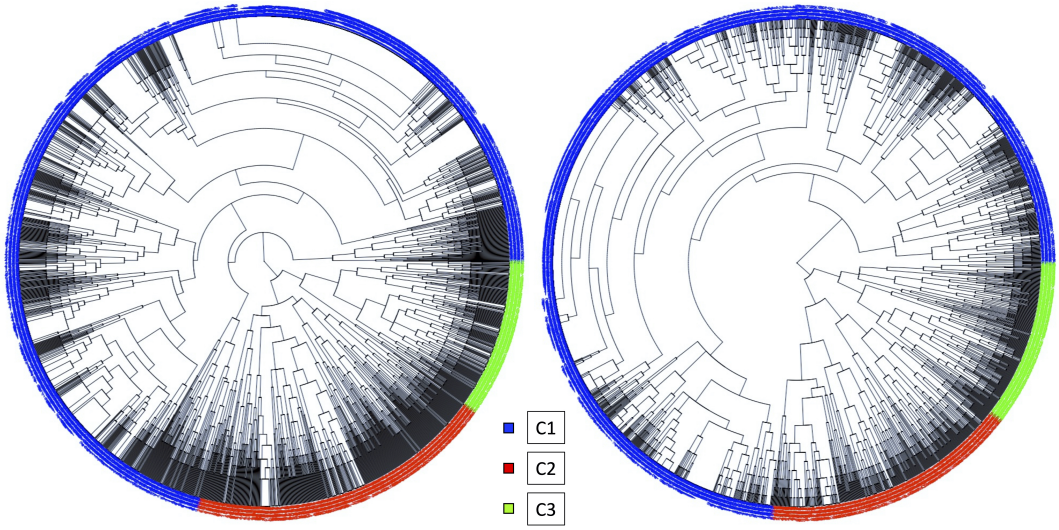

Figure XXa: Circular dendogram with three-category classifications distance-Euclidean
Figure XXb: Circular dendogram with three-category classifications distance-Manhattan

Cluster assignments are formed using the median traditional sum value within each cluster-:

Table XX:

Cluster	Pop L^2	Med. Sum L^2	Pop. L^1	Med. Sum L^1
C1	1768	3	1834	4
C2	246	10	261	13
C3	481	18	400	19

Table XX: Hierarchical clustering with Euclidean (L^2) metric and Manhattan (L^1) assignments based upon median traditional sum value

2.4.3.3 Self Organizing Maps

Self-organizing maps (SOMs) require a pre-specified structure (generally a lattice structure) to which observations are clustered based upon projected proximity [6]. SOMs are not restricted to linear model frameworks, and they preserve topological relationships between variables in the mapping from higher dimensions [6]. SOMs have useful visualization methods that are meaningful and can be used for deduction and inference [7].

In this analysis SOMs are fit to a 15x15 lattice structure with a hexagonal topology and both L^2 and L^1 metrics. Data is standardized prior to completing all SOM clustering procedures. Plots of each of these mappings are found in **Figure XXa** L^2 metric and **Figure XXb** L^1 metric below

Figure XXa and XXb:

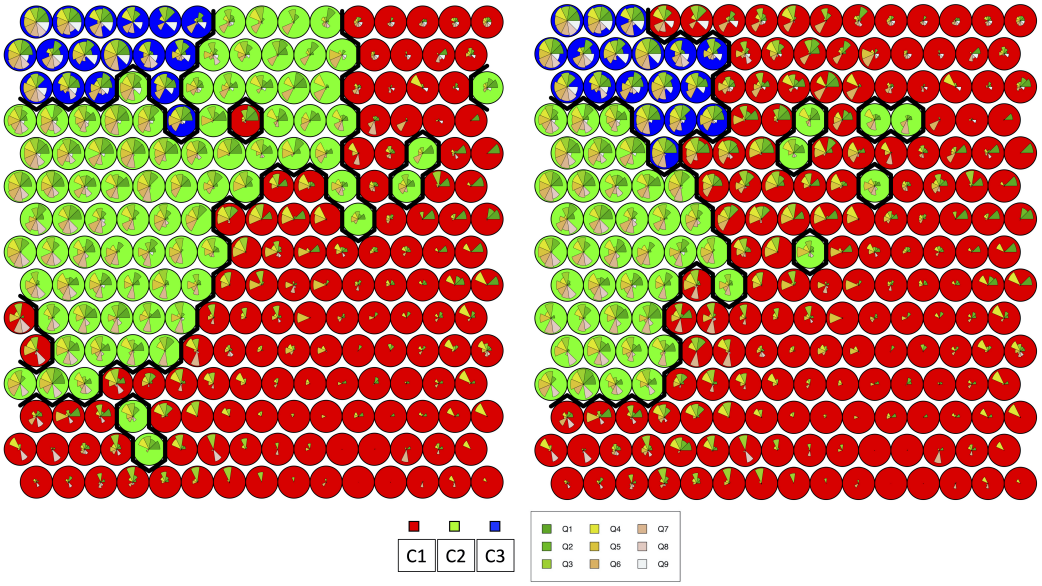

Figure XXa: Self Organizing Map with Euclidean metric, 15x15 hexagonal topology lattice with L^2 metric **Figure XXb**: Self Organizing Map with Manhattan metric, 15x15 hexagonal topology lattice with L^2 metric

Cluster assignments are formed using the median traditional sum value within each cluster-:

Table XX:

Cluster	Pop L^2	Med. Sum L^2	Pop. L^1	Med. Sum L^1
C1	1790	3	2006	4
C2	122	15	155	16
C3	583	21	334	21

Table XX: Self Organizing Map clustering with Euclidean (L^2) metric and Manhattan (L^1) assignments based upon median traditional sum value

3 Analysis and Results

3.1 Inferential Outcome Simulation

Figure XX:

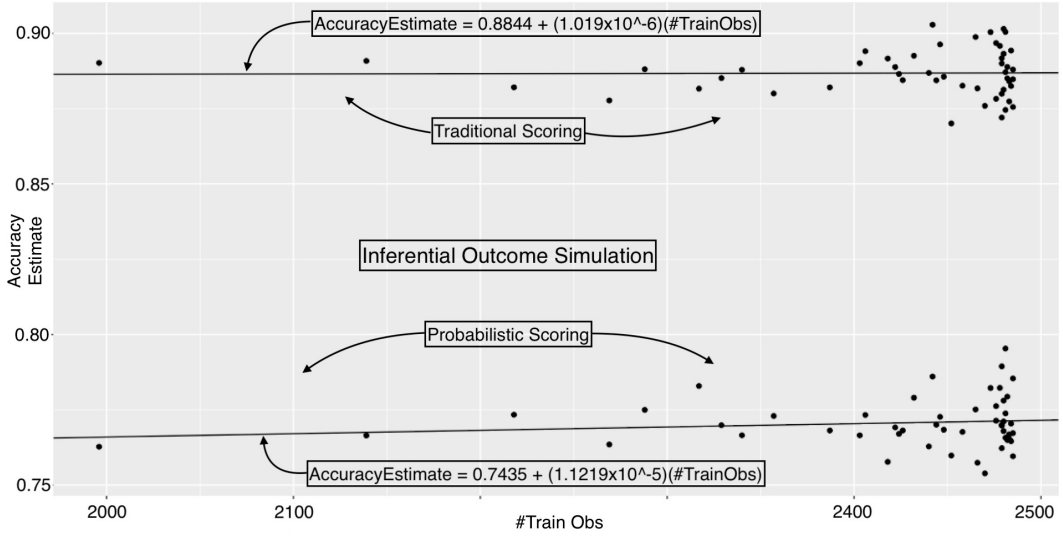

Figure XX: displays estimated accuracy values for probabilistic and traditional classifications across 50 values of training set definitions. Also plotted are Linear Regression trend lines for each estimation method's accuracy estimates.

- **Table XX:**

Parameter	Estimate	Std. Error	t. Value	Pr(>|t|)
Intercept	8.844e-1	2.902e-2	3.048e1	<2e-16
ID = Pscore	-1.409e-1	4.104e-2	-3.433e0	8.840e-4
#Train Obs	1.019e-6	1.197e-5	8.850e-2	9.323e-1
(ID = Pscore):(#Train Obs)	1.020e-5	1.692e-5	6.030e-1	5.481e-1

Table XX: summary table for linear regression trend lines fit to accuracy estimate values displayed in **Figure XX**

After simulating an outcome using the Inferential Simulation method outlined in **Section (XX)** accuracy values of probabilistic and tradition classification are plotted for 50 different

definitions of training sample length. The results of this comparison show that, when measured agains the inferentially simulated outcome, traditional scoring classification is more accurate than probabilistic scoring for all defined training samples. The results indicate that probabilistic classification accuracy grows at a faster rate than traditional classification accuracy; however this result lacks sufficient evidence to support the claim of a significant finding (p=0.5581).

3.2 Probabilistic Simulations

Figure XX:

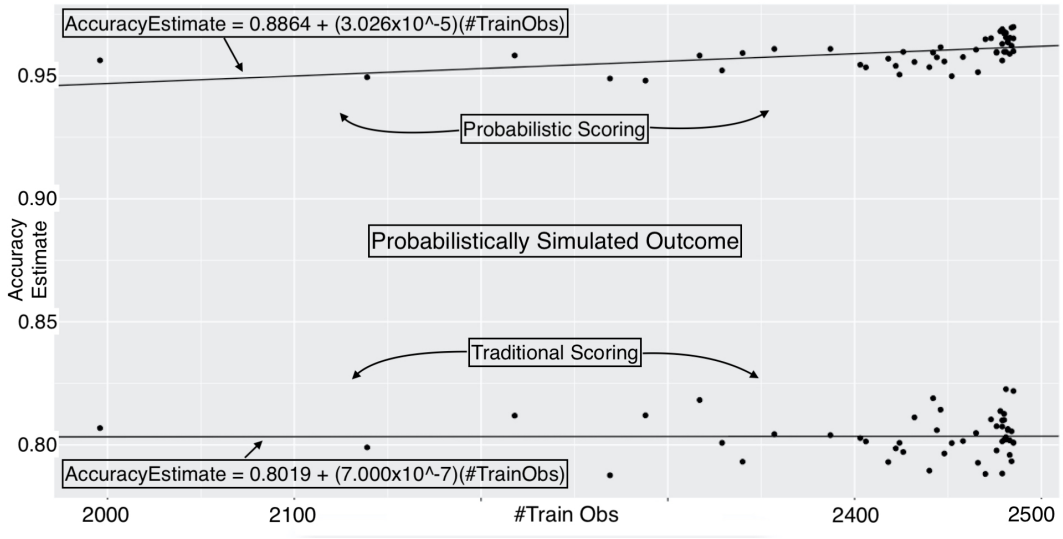

Figure XX: displays estimated accuracy values for probabilistic and traditional classifications as estimated using a probabilistically simulated outcome variable, and calculated across 50 values of training set definitions. Also plotted are Linear Regression trend lines for each estimation method's accuracy estimates.

Table XX:

Parameter	Estimate	Std. Error	t. Value	Pr(>|t|)
Intercept	8.864e-1	2.503e-2	3.541e1	<2e-16
ID = Traditional	-8.453e-2	3.540e-2	-2.388e0	1.891e-2
#Train Obs	3.026e-5	1.032e-5	2.932e0	4.210e-3
(ID = Traditional):(#Train Obs)	-2.956e-5	1.460e-5	-2.025e0	4.562e-2

Table XX: summary table for linear regression trend lines fit to accuracy estimate values displayed in **Figure XX**

After simulating an outcome using the Probabilistic Simulation method outlined in **Section (XX)** accuracy values of probabilistic and tradition classification are plotted for 50 different definitions of training sample length. The results of this calculation show that, when measured against the probabilistically simulated outcome, probabilistic classification is more accurate than traditional classification for all defined training samples. The results indicate that probabilistic scoring accuracy grows at a faster rate than traditional scoring accuracy, and that this effect is supported by evidence in the probabilistically simulated data (p=0.00421)

3.3 Unsupervised Algorithm Outcomes

3.3.1 Kmeans Clustering

Figure XX:

Figure XX: displays estimated accuracy values for probabilistic and traditional classifications as estimated using outcomes resulting from a Kmeans clustering process. The accuracy values are calculated across 50 values of training set definitions. Also plotted are Linear Regression trend lines for each estimation method's accuracy estimates.

- **Table XX:**

Parameter	Estimate	Std. Error	t. Value	Pr(>|t|)
Intercept	5.389e-1	2.264-2	2.380e1	<2e-16
ID = Traditional	-4.129e-1	3.202e-2	-1.290e1	<2e-16
#Train Obs	-1.483e-6	9.321e-6	-1.590e-1	8.740e-1
(ID = Traditional):(#Train Obs)	1.082e-5	1.318e-5	8.210e-1	4.130e-1

Table XX: summary table for linear regression trend lines fit to accuracy estimate values displayed in **Figure XX**

A Kmeans unsupervised learning algorithm was used to generate response simulations to calculate the accuracy of probabilistic and traditional classification methods for 50 different definitions of training sample length. The results of this calculation show that, when measured against responses generated from a Kmeans unsupervised learning algorithm, probabilistic classification is more accurate than traditional classification for all training samples defined.

The results indicate that probabilistic scoring accuracy does not continue to grow as training sample size increases; however, this result lacks sufficient evidence to support the claim of a significant finding (p=0.874).

3.3.2 Hierarchical Clustering

Figure XX:

Figure XX: displays estimated accuracy values for probabilistic and traditional classifications as estimated using outcomes resulting from a hierarchical clustering process. The accuracy values are calculated across 50 values of training set definitions. Also plotted are Linear Regression trend lines for each estimation method's accuracy estimates.

Table XX:

Parameter	Estimate	Std. Error	t. Value	Pr(>|t|)
Intercept	3.437e-1	2.505e-2	1.372e1	<2e-16
ID = Traditional	3.805e-1	3.543e-2	1.074e1	<2e-16
#Train Obs	1.548e-5	1.031e-5	1.501e0	1.350e-1
(ID = Traditional):(#Train Obs)	-9.187e-6	1.459e-5	-6.300e-1	5.300e-1

Table XX Caption: summary table for linear regression trend lines fit to accuracy estimate values displayed in **Figure XX**

A hierarchical clustering learning algorithm was used to generate response simulations to calculate the accuracy of probabilistic and traditional classification methods for 50 different definitions of training sample length. The results of this calculation show that, when measured against responses generated from a hierarchical clustering algorithm, traditional classification is more accurate than probabilistic scoring for all training samples defined. The results indicate that probabilistic classification accuracy increases with training sample size; however, this result lacks sufficient evidence to support the claim of a significant finding (p=0.135).

3.3.3 Self Organizing Maps

Figure XX:

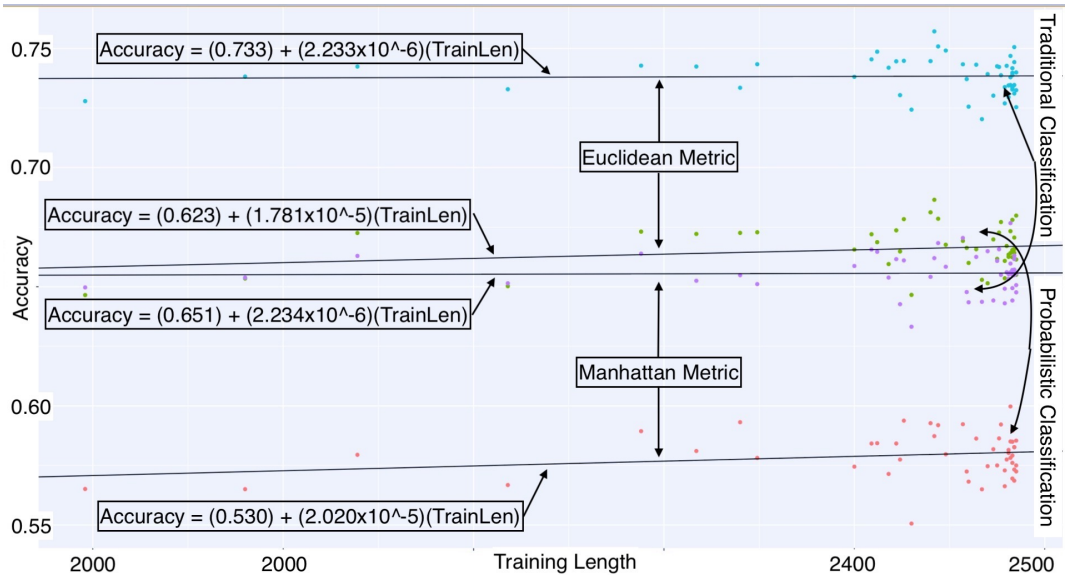

Figure XX: displays estimated accuracy values for probabilistic and traditional classifications obtained using outcomes estimated by two Self Organizing Maps that have a 15x15 lattice structure, a hexagonal topology, and employed either a Euclidean or Manhattan metric. Also plotted are Linear Regression trend lines for each estimation method's accuracy estimates.

Table XX:

Parameter	Estimate	Std. Error	t. Value	Pr(>|t|)
Intercept	6.227e-1	2.814e-2	2.213e1	<2e-16
ID = (Pscore, L^1)	-9.234e-2	3.979e-2	-2.321e0	2.136e-2
ID = (Trad, L^2)	1.102e-1	3.979e-2	2.769e0	6.170e-3
ID = (Trad, L^1)	2.841e-2	3.979e-2	7.140e-1	4.762e-1
#Train Obs	1.781e-5	1.60e-5	1.535e0	1.264e-1
ID = (Pscore, L^1):(#Train Obs)	2.390e-6	1.640e-5	1.460e-1	8.843e-1
ID = (Trad, L^2):(#Train Obs)	-1.557e-5	1.640e-5	-9.490e-1	3.437e-1
ID = (Trad, L^1):(#Train Obs)	-1.593e-5	1.640e-5	-9.710e0	3.329e-1

Table XX: summary table for linear regression trend lines fit to accuracy estimate values displayed in **Figure XX**

Self Organizing Maps were used to simulate response simulations to estimate the accuracy of probabilistic and traditional classification methods for 50 different definitions of training sample length. The results of this calculation show that, when measured against responses generated by a 15x15 hexagonal topology lattice SOM (either Euclidean or Manhattan metrics), traditional classification is more accurate than probabilistic classification for all training samples defined. The results indicate that probabilistic classification accuracy increases with training sample size; however, in both SOM simulations (Euclidean and Manhattan metrics) this result lacks sufficient evidence to support the claim of a significant finding (p=0.1264, L^2) and (p=0.8843, L^1)

4 Discussion

4.1 Summary of Methods

This analysis has performed calculations of probabilistic and traditional classification accuracy using multiple outcomes simulated from information contained in PHQ9 data that should reasonably represent a hierarchy of depression classification categories. It is hoped that, by providing a spectrum of theoretically and empirically reasonable accuracy measurements, this analysis can begin to elucidate those values of accuracy that are likely to represent reality. Six outcome variables were simulated in three different ways:

- Inferential Simulations are based on theoretical information relating to experimentation
- Probabilistic Simulations are also based on theoretical information relating to experimentation, but also incorporate empirical evidence in the data
- Unsupervised Methods do not account for any underlying experimentation practices. These outcomes are generated based upon a hierarchy of relations imposed on the results of a clustering algorithm.

Calculations were evaluated at 50 different definitions of training sample length, and trend lines were fit to these calculated accuracy values to estimate the relationship between accuracy and training sample length.

4.2 Summary of Results

Table XX:

Outcome Simulation Method	Superior Model Accuracy	Min Train Set Size Equality	Max Train Set Size Equality
Inferential	Traditional	1.256e4	1.381e4
Probabilistic	Probabilistic	NA	NA
KMeans	Probabilistic	NA	NA
Hierarchical	Traditional	2.455e4	4.135e4
SOM Euclidean	Traditional	6.176e3	7.106e3
SOM Manhattan	Traditional	5.500e3	6.122e3

Table XX: Analysis results summary. **Table XX** above lists the modeling method found to be more accurate with each corresponding simmulated response. In the cases where probabilistic classification is not the preffered method, **Table XX** also lists a range of estimated training set sizes that, if surpassed, would theoretically allow probabilistic classification accuracy to exceed traditional classification accuracy when measured against the corresponding response.

Training set size ranges in **XX** were obtained by solving for the intersection of trend lines in each simulation method. Minimum range value estimates assume that traditional scoring accuracy trends have zero accuracy growth, and maximum range value estimates assume the estimated positive growth rate in the traditional classification accuracy trend line.

The analysis conducted must conclude without a definitive conclusion concerning the comparison of probabilistic classification accuracy and traditional classification accuracy. In two of the six simulated outcomes evaluated, probabilistic classification accuracy is judged to be superior to traditional classification accuracy. As this answer is not deffinitive, this analysis has failed to produce a directly useable comparison of probabilistic and traditional

classification accuracy.

Conversely, this analysis has produced confirmational evidence that probabilistic classification accuracy increases with training sample size. Excluding only the Kmeans simmulated respose, the probabilistic classification accuracy had a higher positive trend than did traditional scoring classification accuracy. Further research is required in order to provide a complete range of estimated training sample set sizes over which accuracy can be calculated.

The Kmeans simmulated outcome is expected. Intuition would dictate that obtaining more observations should not decrease probabilistic scoring accuracy. However, increasing observational counts will increase initialization dependencies in the Kmeans algorithm. Therefore, the decrease in probabilistic scoring classification accuracy is reasonably explained by a tendency of the outcome variable to "shift its preference" when different training data sizes are provided.

4.3 Client Satisfaction Concerns

As initially specified, this analysis has not completed either of the client-specified goals. Mathematical proof aside, this analysis has failed to produce any concrete comparison of probabilistic and traditional scoring accuracies. The CV analysis has also failed to produce reliable results and needs to be adapted to allow for more flexible sampling practices. Additionaly, the analysis does not address or even acknowledge the client's second stated goal.

4.4 Next Steps

Although the approach taken in this analysis might be interesting, additional simmulations, further or more extensive training sample size evaluations, and other finishing touches are not advisable as they will most likely not produce conclusive results. However, improvements

could be easily implemented by independently sampling Test and Training sets, then replacing observations as sampling takes place. This new process would allow for an increased number of Train/Test sample size pairings.

Self Organizing Maps have improved algorithms which allow for the structure on which they are built to be "optimized". This so called "Growing Self Organizing Map" (GSOM) would better capture information in the original data than pre-specified structures, and would be an appropriate algorithmic improvement.

---

5 Appendix

5.1 Data and Code Availability

Access to code, with instructions, and all data used for the analysis completed here is available for download at:

`https://github.com/leepanter/PScoreVSTscore`

`git@github.com:leepanter/PScoreVSTscore.git`

5.2 List of Figures and Tables

List of Figures

List of Tables

5.3 References

1. Kroenke K, Spitzer RL (2002) The phq-9: A new depression diagnostic and severity measure. *Psychiatric annals* 32: 509–515.

2. Kroenke K, Spitzer R (2010) Instruction manual: Instructions for patient health questionnaire (phq) and gad-7 measures.

3. Kroenke K Spitzer, rl & williams, jb (2001). The phq-9. *Journal of General Internal Medicine* 16: 606–613.

4. Malik A More effective and cost effective use of the phq-9.

5. (2020) *Sthdacom.*

6. Belavkin R Lecture 13: Self-organising maps.

7. Tanner D (2020) Introduction to self organizing maps in r - the kohonen package and nba player statistics. *Githubio.*