2019

Development of a Data-driven Patient Engagement Score Using Finite Mixture Models

Eric Bae
South Dakota State University

DEVELOPMENT OF A DATA-DRIVEN PATIENT ENGAGEMENT SCORE USING FINITE MIXTURE MODELS

BY

ERIC BAE

A thesis submitted in partial fulfillment of the requirements for the

Master of Science

Major in Mathematics

Specialization in Statistics

South Dakota State University

2019

DEVELOPMENT OF A DATA-DRIVEN PATIENT ENGAGEMENT SCORE USING FINITE MIXTURE MODELS

ERIC BAE

This thesis is approved as a creditable and independent investigation by a candidate for the Master of Science in Mathematics with a Specialization in Statistics degree and is acceptable for meeting the thesis requirements for this degree. Acceptance of this does not imply that the conclusions reached by the candidate are necessarily the conclusions of the major department.

Semhar Michael, Ph.D.
Thesis Advisor Date

Kurt Cogswell, Ph.D.
Head, Department of Mathematics & Statistics Date

Dean, Graduate School Date

ACKNOWLEDGMENTS

I would like to thank my advisor, Dr. Semhar Michael, who has assisted me through my graduate studies, and my mentor Dr. Gemechis Djira for constructive feedback and statistical knowledge, which helped with my research immensely. I would also like to thank Dr. Emily Griese, Dr. Benson S. Hsu, and Cheryl Stansbury from Sanford Health Collaborative for providing me Electronic medical records data set and further assistance, which allowed me to complete this research, Dr. Kurt Cogswell and the Mathematics & Statistics Department for supporting me through the financial assistance and providing me the opportunity to work as a Graduate Teaching Assistant. Finally, I would like to acknowledge Dr. Tanya Gupta for volunteering to serve as the Graduate Faculty Representative on my committee.

LIST OF FIGURES

LIST OF TABLES

ABSTRACT

DEVELOPMENT OF A DATA-DRIVEN PATIENT ENGAGEMENT SCORE USING FINITE MIXTURE MODELS

ERIC BAE

2019

Patient activation measure (PAM) is widely adopted by health care providers to access individuals' knowledge, skill, and confidence for managing one's health and healthcare. Patient activation measure (PAM), licensed by Insignia Health, is widely adopted by health care providers to access individuals' knowledge, skill, and confidence for managing one's health and healthcare. Multiple studies corroborate the effectiveness of activation measure in predicting most health behaviors, including preventive behaviors, healthy behaviors, self-management behaviors, and health information seeking. However, PAM is heavily dependent on subjective patient-reported data, which are often incomplete. The purpose of this study is to develop an objective statistical model to create a score derived from patient behavioral measurements. Ranging from 1 to 3, the score, which we named **patient engagement score (PES)**, was derived entirely from three objective variables - number of immunization, number of missed scheduled visits, and rate of patient adherence to prescription refill - using finite mixture model and EM algorithm. Finally, we performed simple and multiple linear regressions for the association between PES and each of the health-related outcomes.

1 INTRODUCTION

Increasing number of studies suggest that patient engagement can lead to better health outcomes, improve quality of care and patient safety, and help control health care costs.[18] The long term goal of health care management is to develop as well as implement various strategies for engaging with patients who do not meet adherence to treatment.

In 2005, Patient Activation Measure® (PAM®) was developed and tested by Hibbard et al. to allow improved assessment of patient's individual competencies for self-assessment.[8] The definition of patient activation is individual skill, knowledge, and confidence in managing their health, and health care.[8] The instrument was initially developed in longer versions of 22 survey items and later adapted to a shorter version with 13 survey items.[7] The shortened 13-item version of PAM® is currently in use because the measure has succeeded in maintaining adequate precision.[7] The PAM® survey measures patients on a 0 – 100 scale, then segment patients into one of four activation levels.[8]

According to Insignia Health, each point increase in PAM® score correlates to a 2 % decrease in hospitalization and 2 % increase in medication adherence, and demographics and socioeconomic factors account for 6 % or less of the variation in PAM® scores.[6] Further, traditional retrospective clinical risk models take years to deliver relevant results, and still fail to identify more than half of patients in the lower two activation levels.[6] As such, PAM® was regarded as a great substitute for predicting health conditions of patients, especially when more advanced measurement models are lacking, and is already widely adopted by health care providers across the United States, as well as a number of European countries.[17] However, modern patient activation measurement technique has its shortfalls: it depends heavily on patients self assessment and requires a survey to be conducted for individual patients.[10] This may tend to be expensive and sometimes patients self assessment may change as time

progresses. Another issue is that PAM® was developed with a sample of predominantly English-speaking European Americans, potentially leading to bias issues and may hinder its adaptation in regions where patients do not share similar health-related behaviors [1].

Besides PAM®, there have been several studies that have attempted to introduce metrics and evaluation tools that measure patient engagement.[18] A study by Dendere et al. (2019) suggests that increased use of inpatient and outpatient portal (IPP) is associated with improved engagement level.[2] Another, more recent study by Macklin et al. (2019), introduces Engaging Patients in Care (EPIC), a local patient engagement initiative at University Health Network (UNH) for patients and families who have received care for heart failure (HF), heart transplant (HT), or mechanical circulatory support (LVAD) throughout Ontario, Canada.[11] Under EPIC, patients will be separated into four levels of engagement with its assessment based on four engagement priorities - care delivery/policy, patient advocacy, research, and paper support work while removing direct care from consideration.[11] While promising, at the time of the writing, EPIC was in early stages of development. The team is expected to pilot EPIC with a small group of HT recipients at Toronto General Hospital and get them involved through orientation and training, create a terms of reference, use the online bulletin board, and complete the specific task of updating the hospital's HT manual with their lived experience within the Care Delivery/Policy engagement priority.[11]

Our objectives in development of effective patient engagement methods include measuring the level of the patient engagement by developing a statistical model and evaluating the association between patient engagement level and health outcomes. By this goal, we expect it can be helpful for the hospital or clinic teams to understand the patient's behaviors, so they can implement patient-specific interventions and make patients more confident to manage their health. The aim of this project is measuring the level of the patient engage-

ment using a small number of readily-available behavioral variables existing in medical records. This differs from other patient engagement metrics and tools in use, such as PAM® and EPIC, both of which rely more heavily on subjective outputs and active involvement by health care providers. Our model considers the engagement levels as latent or hidden variables and the observed patient behavior variables as input. After creating the model, we will assign the Patient Engagement Score (PES) to patients, which can help health care facilities to better understand their patients using latent variable modeling scheme through finite mixture models.

In 2018, Ngorsuraches et al. have attempted to develop a preliminary score, also referred to as PES, using patient health outcome measures from a multivariate Gaussian mixture model.[14] However, the score generated some contradicting results, with some cases indicating higher engagement score to be associated with worsened health-related outcomes.[14] In this thesis, we developed mixture models of multivariate distribution and we specify the PES by ramifying some methodologies adopted by Ngorsuraches et al. The rest of the thesis is structured as follows: in Section 2, we introduce finite mixture models, parameter estimation, model selection and regression; in Section 3, we present the data analysis including exploration, model fitting and selection, and results; in Section 4, we conclude with a discussion, where we summarize and compare our results to the results obtained by the preliminary score developed by Ngorsuraches et al. The derivation of the mixture model parameter estimation and additional findings not included in the main paper are given in the Appendix (Section A).

2 METHODOLOGY

2.1 THE PROBABILISTIC MODEL

Due to the nature of the patient engagement score, application of finite mixture models and the Expectation-Maximization (EM) algorithm is the central focus of this paper. Therefore the preliminary discussion on both is given below.

In a finite mixture model, the data are assumed to have arisen from a mixture of an initially specified number of populations in different proportions.[16] Suppose we have a random sample $\boldsymbol{Y} = \{\boldsymbol{Y}_1, ..., \boldsymbol{Y}_N\}$ with size N, and $\boldsymbol{Y}_i$ be a D-dimensional random vector with a mixture density function $g(\boldsymbol{y}_i)$. Then the mixture density function of $\boldsymbol{Y}_i$, can be written in the form

$$g(\boldsymbol{y}_i; \boldsymbol{\psi}) = \sum_{k=1}^{K} \pi_k f(\boldsymbol{y}_i; \boldsymbol{\theta}_k), \tag{2.1}$$

where $f(\boldsymbol{y}_i; \boldsymbol{\theta}_k)$ is the component density of the mixture, $\boldsymbol{\theta}_k$ is the true component-specific parameter vector of a density function f and depends on the type of distributions of our sample, and π_k is the prior probability of the k-th component (alternatively referred to as weight or mixing proportion).[16] The mixing proportions have the following constraints: $0 < \boldsymbol{\pi}_k \leq 1$ and $\sum_{k=1}^{K} \pi_k = 1$. Finally, $\boldsymbol{\psi}$ is a vector of all unknown parameters of interests, $(\pi_1, ..., \pi_{K-1}, \boldsymbol{\theta}'_1, ..., \boldsymbol{\theta}'_K)'$.[16] Note that $\pi_K \notin \boldsymbol{\psi}$ because π_K can be easily obtained as $\pi_K = 1 - \sum_{k=1}^{K-1} \pi_k$. The log-likelihood function for independent and identically distributed (iid) observations from $g(\boldsymbol{y_i}; \boldsymbol{\psi})$ with a sample size of N is written in the form

$$\log L(\boldsymbol{\psi}|\boldsymbol{y}_i) = \sum_{i=1}^{N} \log \sum_{k=1}^{K} \pi_k f(\boldsymbol{y}_i; \boldsymbol{\theta}_k). \tag{2.2}$$

The log-likelihood function, Equation 2.2, is obtained by taking the product of Equation 2.1 N times, then calculate the natural logarithm of the product. In many real world applications, the component distributions $f(\boldsymbol{y}_i; \boldsymbol{\theta}_k)$ are unknown; hence, the parameter $\boldsymbol{\theta_k}$ will take different forms depending on the

distributions we choose to apply.[5]

If $\boldsymbol{y}_i = (y_1, y_2, \ldots, y_D)'$ is D-dimensional and it is assumed that y_d's are mutually independent of one another, then the mixture density in Equation 2.1 can be written as

$$g(\boldsymbol{y}_i; \boldsymbol{\psi}) = \sum_{k=1}^{K} \pi_k \prod_{d=1}^{D} f_d(y_{id}; \boldsymbol{\theta}_{kd}), \tag{2.3}$$

where the joint density, $f(\boldsymbol{y}_i; \boldsymbol{\theta}_k)$ from Equation 2.1, is the product of marginal densities, and each list elements can have different set of dependent variable and correspond to f_d. In this paper, independence were assumed and the data was modeled by a mixture of three of different distributions - Negative Binomial, Gamma, and Gaussian. Gamma and Negative Binomial distributions was used because their domains ($\in (0, \infty)$) roughly correspond to our observations. Further, Negative binomial was substituted for Poisson because the former outperformed Poisson in observations with varying dispersion. Gaussian was included because it is one of the most commonly applied distribution in mixture models.

As an example, one of the mixture models tested, namely, Gaussian, Negative Binomial, and Gamma, would result in the following density function:

$$g(\boldsymbol{y}_i; \boldsymbol{\psi}) = \sum_{k=1}^{K} \pi_k f_{Gaus}(y_{i1}; \mu_k, \sigma_k^2) f_{NB}(y_{i2}; r_k, p_k) f_{Gam}(y_{i3}; \alpha_k, \beta_k), \tag{2.4}$$

where $r_k, p_k, \mu_k, \sigma_k^2, \alpha_k, \beta_k$ represent true parameters of interest for the corresponding distributions. The log-likelihood function of the model in Equation 2.4 would be defined as

$$\log L(\boldsymbol{\psi}|\boldsymbol{y}_i) = \sum_{i=1}^{N} \log \sum_{k=1}^{K} \pi_k f_{Gaus}(y_{i1}; \mu_k, \sigma_k^2) f_{NB}(y_{i2}; r_k, p_k) f_{Gam}(y_{i3}; \alpha_k, \beta_k), \tag{2.5}$$

where

$$f_{Gaus}(y_{i1};\mu_k,\sigma_k^2) = \frac{1}{\sigma\sqrt{2\pi}} e^{-(y_{i1}-\mu_k)^2/2\sigma_k^2}, \tag{2.6}$$

$$f_{NB}(y_{i2};r_k,p_k) = \frac{\Gamma(y_{i2}-r_k)}{\Gamma(r_k)\Gamma(y_{i2}+1)} p_k^{r_2}(1-p_k)^{y_{i2}}, \text{ and} \tag{2.7}$$

$$f_{Gam}(y_{i3};\alpha_k,\beta_k) = \frac{{y_{i3}}^{\alpha_k-1} e^{-\left(y_{i3}/\beta_k\right)}}{{\beta_k}^{\alpha_k}\Gamma\left(\alpha_k\right)}. \tag{2.8}$$

Crucially, the three variables - y_{i1}, y_{i2}, and y_{i3} - are assumed to be independent of one another. The three density functions listed - Equations 2.6, 2.7, and 2.8 - depend on different parameters, which require maximum likelihood estimation to obtain the estimates.

2.2 ESTIMATION OF PARAMETERS

Suppose component memberships of all observations were known. Then we can introduce a random variable $\boldsymbol{Z} = \{Z_1, ..., Z_N\} \in \{1, ..., K\}$ where each Z_i represents the true component membership of the observation y_i. Since no observations are assumed to belong to more than one class, we can say that the likelihood of an observation at a class other than its own would be non-existent. Therefore, the likelihood and the log-likelihood function of the complete data become

$$\begin{aligned} L_c(\boldsymbol{\psi}|\boldsymbol{y}_i) &= \prod_{i=1}^{N}\prod_{k=1}^{K}\pi_k \prod_{d=1}^{D} f_d(\boldsymbol{y}_{id};\boldsymbol{\theta}_{kd})^{I(z_i=k)}, \\ l_c(\boldsymbol{\psi}|\boldsymbol{y}_i) &= \sum_{i=1}^{N}\sum_{k=1}^{K} I(z_i=k)\sum_{d=1}^{D}\log(\pi_k f_d(\boldsymbol{y}_{id};\boldsymbol{\theta}_{kd})), \end{aligned} \tag{2.9}$$

where $I(z_i = k)$ is an indicator variable $\in \{0, 1\}$. For k in which an observation y_i belongs, $I(z_i = k) = 1$ and $I(z_i \neq k) = 0$ for the k in which it does not. This ensures that $\sum_{d=1}^{D} I(z_i \neq k)\log(\pi_k f_d(\boldsymbol{y}_{id};\boldsymbol{\theta}_{kd})) = 0$, leaving just the likelihood of the density function where $z_i = k$. Calculating maximum likelihood estimates becomes vastly simpler in this case. However, in this

study, component memberships of the observations are unknown, rendering it incomplete. As such, Expectation-Maximization (EM) algorithm is necessary to compute maximum likelihood estimates of the parameters of interest. EM algorithm simplifies computation of maximum likelihood estimates by linking complete-data model with the incomplete observed structure.[16]

The EM algorithm consists of two steps: Expectation (E-step) and Maximization (M-step). The process of estimating the conditional expectation of the complete-data log-likelihood function is referred to as E-step. In E-step, we define function $Q(\boldsymbol{\psi}|\boldsymbol{\psi}^{(s)})$ as the expectation of the complete-data log-likelihood of $\boldsymbol{\psi}$ given Z and Y. $Q(\boldsymbol{\psi}|\boldsymbol{\psi}^{(s)}) = \mathbb{E}_{\boldsymbol{Z}|\boldsymbol{Y},\boldsymbol{\psi}^{(s)}}[\log L_c(\boldsymbol{\psi}|\boldsymbol{Y},\boldsymbol{Z})]$. The E-step will yield the posterior probabilities of each component. In M-step we maximize the $Q(\boldsymbol{\psi}|\boldsymbol{\psi}^{(s)})$ function subject to the restriction that $\sum_{k=1}^{K}\pi_k = 1$. The steps are iterated until a pre-specified membership convergence criterion is achieved.[3] Assuming that a latent variable $z_i \in \{1, ..., K\}$ exists for each observation $\boldsymbol{y}_i$ and the number of component membership exists for each observation k. If $\boldsymbol{y}_i$ comes from component k, then $z_i = k$. The z_i are the unobserved component memberships and are treated as missing values and the data is augmented by the estimates of component memberships. The estimated a-posterior probabilities denoted as $\widehat{\tau}_{ik}$. For a sample of N observation, $\{(y_1), \ldots, (y_N)\}$, the EM estimator is given by:

E-step: given $\boldsymbol{\psi}^{(s)}$ is the current parameter estimate in the s^{th} iteration, replace the missing data by the estimated a-posteriori probabilities:

$$\tau_{ik}^{(s+1)} = \frac{\pi_k^{(s)} f(\boldsymbol{y_i}; \boldsymbol{\theta}_k^{(s)})}{\sum_{k=1}^{K} \pi_k^{(s)} f(\boldsymbol{y_i}; \boldsymbol{\theta}_k^{(s)})}. \tag{2.10}$$

M-step: given $\tau_{ik}^{(s+1)}$ as the a-posteriori probabilities, we obtain new estimates $\boldsymbol{\psi}^{(s+1)}$ of the parameter by maximizing:

$$Q(\boldsymbol{\psi}^{(s+1)}|\boldsymbol{\psi}^{(s)}) = Q_1(\boldsymbol{\theta}^{(s+1)}|\boldsymbol{\psi}^{(s)}) + Q_2(\boldsymbol{\phi}^{(s+1)}|\boldsymbol{\psi}^{(s)}), \tag{2.11}$$

where

$$Q_1(\boldsymbol{\theta}^{(s+1)}|\boldsymbol{\psi}^{(s)}) = \sum_{i=1}^{N}\sum_{k=1}^{K} \tau_{ik}^{(s+1)} \log f(\boldsymbol{y}_i; \boldsymbol{\theta}_k^{(s+1)}); \tag{2.12}$$

$$Q_2(\boldsymbol{\phi}^{(s+1)}|\boldsymbol{\psi}^{(s)}) = \sum_{i=1}^{N}\sum_{k=1}^{K} \tau_{ik}^{(s+1)} \log(\pi_k^{(s+1)}). \tag{2.13}$$

Q_1 and Q_2 can be maximized separately because Q_1 does not contain $\boldsymbol{\phi}^{(s+1)}$ and Q_2 does not contain $\boldsymbol{\theta}^{(s+1)}$. Therefore $\boldsymbol{\theta}^{(s+1)}$ is given by maximization of Q_1, and $\boldsymbol{\phi}^{(s+1)}$ given by maximization of Q_2.

Recall the mixture model, Gaussian, Negative Binomial and Gamma, from earlier, represented by Equations 2.4 and 2.5. Expanding Q_1 and Q_2 under this example will appear as follows:

$$\begin{aligned} Q_1(\boldsymbol{\theta}^{(s+1)}|\boldsymbol{\psi}^{(s)}) = \sum_{i=1}^{N}\sum_{k=1}^{K} \widehat{\tau}_{ik}^{(s+1)} \Big[& \log(f_{Gaus}(y_{i1}; \mu_{k1}^{(s+1)} \sigma_{k1}^{2(s+1)})) \\ & + \log(f_{NB}(y_{i2}; r_{k2}^{(s+1)} p_{k2}^{(s+1)})) \\ & + \log(f_{Gam}(y_{i3}; \alpha_{k3}^{(s+1)} \beta_{k3}^{(s+1)}))\Big], \end{aligned} \tag{2.14}$$

$$Q_2(\boldsymbol{\phi}^{(s+1)}|\boldsymbol{\psi}^{(s)}) = \sum_{i=1}^{N}\sum_{k=1}^{K} \widehat{\tau}_{ik}^{(s+1)} \log(\pi_k^{(s+1)}). \tag{2.15}$$

Details of the maximization process can be found in Section A.

2.3 RANDOM INITIALIZATION

The EM algorithm requires that initial values are defined for all parameters, such as $\pi_k^{(0)}$ and $\psi^{(0)}$. The common way to initiate the EM algorithm is initializing it from a random position.[12] Usually this random initial position is obtained by drawing at random component means in the data set. An extension of the procedure, referred to as *RndEM*, is repeating it M times from different random positions and selecting the solution with the maximum log-likelihood.[12] In this study, we applied *RndEM* to obtain initial values. The steps of the *RndEM* are as followed:

1. Randomly select K observations from $\boldsymbol{y}_i$, $i \in 1, ..., N$. We will denote the selected observations as $\boldsymbol{y}_j$, where $j \in 1, ..., K$.
2. Assign observations to center based on Euclidean distance. The Euclidean distance is defined as

$$d(\boldsymbol{y}_i, \boldsymbol{y}_j) = ||\boldsymbol{y}_i - \boldsymbol{y}_j||_2 = \sqrt{(\boldsymbol{y}_i - \boldsymbol{y}_j)^2}, \tag{2.16}$$

   where $i \in 1, ..., N$, $j \in 1, ..., K$. This will result in K different Euclidean distances for each observation. Assign each observation a value between 1 to K, representing the group to which the observation has the smallest $d(\boldsymbol{y}_i, \boldsymbol{y}_j)$.
3. From the groupings, calculate parameter estimates $(\pi_k^{(0)}, \psi_k^{(0)})_{k=1}^K$.
4. Calculate the log-likelihood based on parameters estimated from step 3.
5. Repeat steps 1 to 4, M number of times. This M needs to be a reasonably large number to minimize fluctuation for repeated experiment.
6. Select the initial parameters based on the highest likelihood value.

This method is known to work well if small K values are considered.[13] Other methods are recommended for large K. In our application, we consider $K = 1, ..., 5$ to obtain interpretable results, hence this initialization method was implemented.

2.4 MODEL SELECTION

In the EM algorithm, there exists model selection problem in determining the number of mixture components K. The problem of choosing the number of the components with the underlying probability model can be reformulated as a statistical model choice problem. Since the classical likelihood ratio test does not hold for mixtures, testing for the number of component is commonly carried out using information criterion. Criteria based on penalized likelihood

such as Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC) have been applied successively when selecting the model.[20] In general, information criteria take the following form:

$$\text{Information Criterion} = J\log(N) - 2\log\hat{L}(\boldsymbol{\psi}|\boldsymbol{y}_i), \tag{2.17}$$

where J is 2 for the AIC and the number of parameters in the model for the BIC. In our example, the number of parameters would be equivalent to $2DK + K - 1$, where 2 represents the number of parameters per distribution (each of the three distributions depend on two distinct parameters), D the total number of variables, K the class membership size.[20] $K - 1$ represents the number of π_k parameters to estimate. Since models with low AIC and BIC are preferable, with the same log-likelihood and sample size, BIC has higher penalty and tend to prefer lower K compared to AIC. The log-likelihood increases as K increases, and improvements in $\log\hat{L}$ will be compared with the increase in penalty suffered from rising K, which ultimately results in rising J as well. In our experiment, we relied on BIC to determine the optimal K.

2.5 MODEL-BASED CLUSTERING

After selecting the mixture model with specific component and variable distributions and number of classes that returns the minimum BIC values, we can apply the a-posteriori probabilities - $\tau_{ik}^{(s+1)}$ - to assign each observation to a class. This is achieved through maximum a-posteriori (MAP) estimation - assigning each observation to the cluster with the largest $\tau_{ik}^{(s+1)}$, thereby creating an output, which we will denote as as $\widehat{z}_i$, where $\widehat{z}_i \in \{1, ..., K\}$.

Our goal is to transform our assigned cluster values, $\widehat{z}_i$, into an ordinal measurement - patient engagement scores. However, the EM algorithm does not automatically assign each observations to clusters in an ordinal fashion. Furthermore, in our analysis, we found that the clusters do not align perfectly

across our three predictors. As demonstrated in Section 4, if we were to realign our clusters so that the mean values of *Immunization* for the classes are in ascending order ($k = 1$ representing the cluster with the lowest average number of *Immunization* and ($k = K$ representing that with the highest), the clusters appear to be out of order in the other two predictors. The same phenomenon occurs if we were to realign them based on the other two variables.

One method we used to reassign cluster membership is to examine interclass and intraclass variances for each predictor, and select the one with the highest intraclass variance and lowest interclass variance. This is achieved by using the following formula for intraclass variance:

$$\boldsymbol{\sigma}_d^2(intra) = \boldsymbol{\sigma}_d^2(total) - \boldsymbol{\sigma}_d^2(inter), \tag{2.18}$$

where $\boldsymbol{\sigma}_d^2(intra) = \{\boldsymbol{\sigma}_1^2(intra), ..., \boldsymbol{\sigma}_D^2(intra)\}$. $\boldsymbol{\sigma}_d^2(total)$ represents total variance for each predictors - *Immunization*, *NoShow*, and *PercLate* - scaled so that they are all 1. $\boldsymbol{\sigma}_d^2(inter)$ represents interclass variance for the aforementioned predictors, and has the form of:

$$\boldsymbol{\sigma}_d^2(inter) = \frac{1}{N-1}\sum_{i=1}^{N}(\boldsymbol{y}_{ikd}^* - \bar{\boldsymbol{y}}_{kd}^*)^2, \tag{2.19}$$

where $\boldsymbol{y}_{ikd}^*$ is a standardized ith observation so that $E[\boldsymbol{y}_{ikd}^*] = 0$ and $Var[\boldsymbol{y}_{ikd}^*] = 1$, and $\bar{\boldsymbol{y}}_{kd}^*$ is the sample mean of the standardized observations for each class k. The process will be discussed in more detail in the Appendix.

2.6 REGRESSION ANALYSIS

After PES scores are assigned to each observation, we performed regression analyses. Multiple logistic regression analysis were performed for association between PES and each of the eight health-related outcomes. These were *Systolic BP, Diastolic BP, HDL, LDL, ED Visit, A1C level, eGFR rate*, and *Hospitalization*. Several patient characteristics and health behavior, namely *Age,*

Tobacco Use, Alcohol Use, Marital Status, Race, Gender, Primary Insurance Type, and *Number of Chronic Conditions*, were added as control variables. Detailed information about the variables are found later in this section. A multiple logistic regression takes the following form:

$$log\left(\frac{\widehat{\pi_i}}{1-\widehat{\pi_i}}\right) = \widehat{\beta_0} + \widehat{\beta_1} PES1_i + \widehat{\beta_2} PES3_i + \widehat{\boldsymbol{\beta_3}}\boldsymbol{X}_i, \tag{2.20}$$

where $\widehat{\pi_i}$ is the predicted probability of a select patient to be in the affirmative for one of the eight aforementioned health-related outcomes. $\widehat{\beta_0}$ represents the estimated intercept of the model. $\widehat{\beta_1}$ and $\widehat{\beta_2}$ represent estimated slope values for observations that were assigned to PES 1 and PES 3, respectively. A correct interpretation of the slope coefficients is given as the change in log-odds of a patient experiencing or having experienced a health-related outcomes for a patient assigned in the given PES compared to the index PES, which is 2. Finally, $\widehat{\boldsymbol{\beta}_3}$ represent a vector of slope coefficients of variables representing patient characteristics and health behavior, and $\boldsymbol{X}_i$ is a matrix of values representing patient characteristics and health behavior. All but two, *Age* and *Number of chronic conditions*, are indicator measurements and all observations will have either 0 or 1 as its values. Note that, in this particular example, there were only three PES classifications. Additionally the slope coefficient for PES 2 does not exist because PES is an ordinal variable obtained through EM algorithm, and PES of 2 is meant to be the "average" observations. PES 2 was selected as the reference point to ensure that the slope parameters, $\widehat{\beta_1}$ of PES 1 and $\widehat{\beta_3}$ of PES 3, would be with respect to PES 2, which is just one unit apart from either. Had we chosen PES 1 and 3 as the index, one of the slope parameters would represent the difference between PES that are two units apart. The result was compared to that between PES and the same outcomes.

Patient-level data was obtained from the Sanford Health electronic medical records along with the PAM scores of a small subset of patients over three

years (2014-2016). The data set contained a total of 147,687 observations with a mix of categorical, such as *Gender*, *Race*, and *Marital Status*, and numerical measurements, including *Age*, *Median Income*, and *Immunization*, for each observation. All outliers, defined as any observations outside the boundary of $[\boldsymbol{y}_i^{median} \pm 1.5 * IQR]$, had been removed. The patients included in our data set were known to suffer from two or more of the following chronic diseases: Diabetes Mellitus, Heart Disease, Kidney Disease, High Blood Pressure and High Cholesterol.

2.7 VARIABLE DEFINITION

The models were developed from the observations whose information was complete. After observations with incomplete data (missing relevant information) were removed, the sample size of the final data set was 4,841.

Four patient-engagement related measurements in our data set were selected as potential indirect measures of patient engagement levels. Some of the measurements were incorporated to generate PES. One, *PercLate*, was derived from the data set. The four measurements were the following:

- *Immunization*: Number of immunizations, such as vaccination, received for the duration of the record time;
- *NoShow*: Number of no-shows at appointed visits in 3 years;
- *Wellness*: Number of wellness exams in 3 years; and
- *PercLate*: This is the measurement that was calculated through examination of dates of prescription and refill for the patients.

The data set also contained the following measurements. These are health-related outcome measurements that were predicted using PES:

- *HDL_Group:* High density lipoprotein groups. This is a binary measurement, where patients were assigned to two groups - Normal ($\geq$ 40mg/dl) and Dyslipidemia ($<$ 40mg/dl). HDL is generally considered a positive

health attribute; therefore individuals with higher HDL is considered to possess superior health outcome than those with lower HDL.

- *LDL_Group:* Low density lipoprotein groups. This is also a binary measurement, where patients were assigned to two groups - Normal (< 100mg/dl) and Dyslipidemia (≥ 100mg/dl). Contrast to HDL, LDL is generally considered a negative health attribute; therefore individuals with lower LDL is considered to possess superior health outcome than those with higher HDL.

- *EDVisit_Group:* Emergency department visits groups. Patients with emergency department visit will be assigned a value of 1 and those without a value of 0, thereby generating a binary $\{0, 1\}$ measurement. 0 is a preferable outcome for patients.

- *A1C_Group:* Glycated hemoglobin (A1C) groups. Patients with A1C level of less than 6.5 % were assigned to "Normal" group and those with greater than or equal to 6.5 % were assigned to "Diabetes" group. "Normal" is the preferred outcome for patients, and was assigned value of 0, whereas "Diabetes" group were assigned to 1.

- *eFGR_Group:* Renal function based on eFGR. The original data set contained the following five groups: "Normal", "Mild", "Moderate", "Severe", and "Kidney failure". However, due to relatively small number of patients belonging to the latter four, they were merged into one massive group, "Not Normal", thereby creating a binary of "Normal" and "Not Normal". "Normal" is considered the more desirable outcome than "Not Normal" and thus was assigned 0.

- *Hospitalization_Group:* This is a binary measurement, where patients without hospitalization within the three year time frame were grouped to "0" and those with at least one hospitalization were grouped to "1".

- *SBP_Group:* Systolic blood pressure groups. "Normal" represents pa-

tients with normal level of systolic blood pressure (< 140mmHg) and "Hypertension" represents those whose SBP was above 140mmHg.

- *DBP_Group:* Diastolic blood pressure groups. "Normal" represents patients with normal level of diastolic blood pressure (< 90mmHg) and "Hypertension" represents those whose DBP was above 90mmHg.

Besides PES, existing measurements representing patient characteristics and health behavior were added as predictors. Their purpose is to represent as our "control variables." These predictors were:

- *Age:* Age in years of patients at the start of the three-year duration. This is one of the two non-binary predictors used in our model.
- *Tobacco:* Whether a patient is a tobacco/cigarette smoker or not. Former smokers were grouped with non-smokers.
- *Alcohol:* Whether a patient is an alcohol consumer or not.
- *Marital:* Marital status of a patient. All non-married individuals - such as single, widowed, and divorced - were grouped as "Single & Others", creating a binary of "Married" and "Single & Others".
- *Race:* Racial heritage of a patient. Due to small sample size of patients who identify as belonging to a race other than white, all non-white patients were grouped into "Non-white" category.
- *Gender:* Gender of a patient. Two genders: "Male" and "Female" were observed.
- *PrimaryInsurance:* Type of primary health insurance of a patient. The patients were divided into two groups - "GOVT" and "Non_GOVT", where the former represents individuals whose coverage is primarily by a public institution, such as Medicare and the latter represents those whose coverage is primarily by private health care providers.

- *Chronic Condition:* Number of chronic conditions suffered by each patient. This is the one of the two predictors that were not binary.

The measurements will be explored further at the subsequent section, Section 3.

3 DATA ANALYSIS

3.1 DATA EXPLORATION

Table 1: Summary of the four measurements selected as indirect measures of patient engagement.

Measure.	**Min.**	**1st Qrt.**	**Median**	**3rd Qrt.**	**Max.**	**Mean**	**SD**	**Size**
Immun	0	2	5	7	37	5.385	4.726	146,302
NoShow	0	0	1	2	17	1.712	2.804	146,145
Wellness	0	0	0	0	3	0.002	0.061	147,651
PercLate	0	0	0	3.101	100	2.845	7.021	12,841

Immunization is an integer-valued variable and ranges from 0 to 37, *NoShow* is an integer variable and ranges from 0 to 17, *Wellness* is an integer value that ranges from 0 to 3, and *PercLate* takes values of any real numbers between 0 and 100, which represents percentage. Figure 1 illustrates distributions of the four measurements.

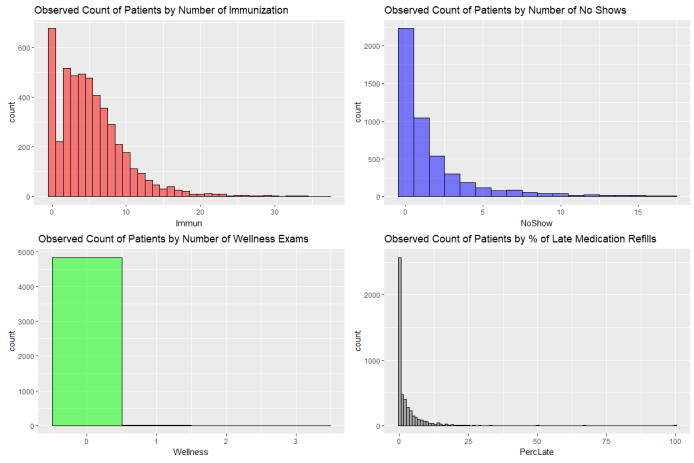

Figure 1: Histograms of the four measurements. $N = 4,841$

All four measurements are skewed heavily to the right and appear to follow

either a zero-inflated distributions or mixtures of distributions. Particularly striking is that *Wellness* appears to experience an extreme case of zero-inflation. We observed only 1,221 across the entire data set and only 7 out of the 4,841 extracted with values greater zero. Therefore, we determined that *Wellness* did not provide meaningful enough information and was subsequently dropped from consideration for our mixture model, leaving with just *Immunization*, *NoShow*, and *PercLate*.

As the exact type of distributions of each measurement is unknown, three distribution types, **Gamma**, **Gaussian (Normal)**, and **Negative Binomial**, are fitted. The number of mixture components K is varied between 1 to 5. Another commonly-applied distribution **Poisson** was not selected because the negative binomial was more flexible in measuring dispersions.

As mentioned in Section 2, assumption of independence among observations and measurements is paramount to the reliance of the calculated log-likelihood, $\log L(\boldsymbol{\psi}|\boldsymbol{y}_i)$. Figure 2 shows the pairwise plots and the linear correlation of the three measurements, and Table 2 presents the significant test results.

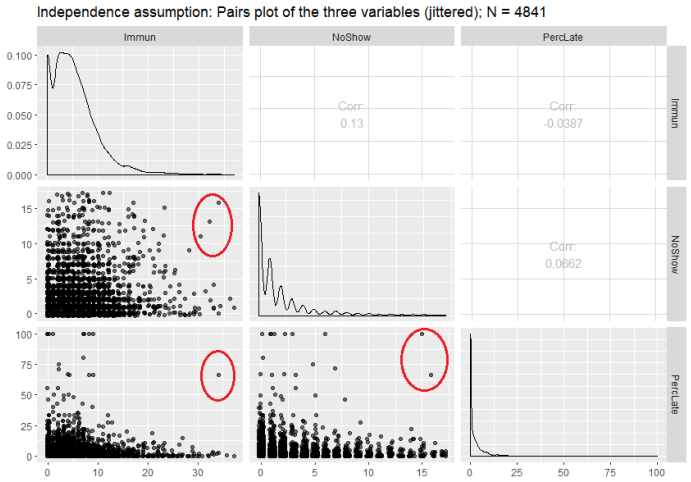

Figure 2: Pairs plots of the three measurements. $N = 4,841$

Table 2: Linear correlations and p-value estimates

Comparisons	Correlation	t-statitic	P-value
Immun vs. NoShow	0.1298	14.6980	$<$ **0.001**
Immun vs. PercLate	-0.0386	-3.7468	**0.007**
NoShow vs. PercLate	0.0662	5.9709	$<$ **0.001**

A quick visual test on Figure 2 already suggests that there may be significant relationship among the variables. Furthermore, each pair of the variables seems to show non-linear relationship but with possible heavy influence from a small number of outliers, circled in red. The calculated linear correlation of each pair of the measurements was between -0.13 and 0.13, which indicate a small correlation among the measurements. However, the extremely-low p-values suggest that the correlations may be significantly different from zero, or that there may be linear relationship among the measurements. This is not surprising as they are selected as proxy for patient engagements, and if they do represent the best predictors for the patient engagement levels, then they should necessarily be related. One may argue that despite low p-values, assumptions of independence are reasonable due to the estimated correlations being relatively low; however, the ultimate conclusion will be left to the readers.

A separate significance test for each PES was performed as well. Output of these tests are found in the Appendix.

3.2 LOG-LIKELIHOOD AND BAYESIAN INFORMATION CRITERIA

Using the $4,841$ observations, we performed EM algorithm using every three permutations of the three distributions - **Gamma**, **Gaussian (Normal)**, and **Negative Binomial**. When initializing, $M = 50$ were picked for every permutation and for every K. The size of each sample was identical to the K that was being tested, meaning for $K = 2$, two observations were randomly selected, assigned into two different classes (K), and every other observation was

assigned to the one of the two classes to which it had the smallest Euclidean distance, from which we calculate log-likelihood based on the permutation of distributions. We repeated this 50 times to obtain 50 different log-likelihoods, and of those, the sample with the highest log-likelihood was chosen to be used to calculate optimal parameter estimates and the clusters by the application of EM algorithm.

Table 2 illustrates log-likelihoods for every K between 1 and 5 inclusive and every three permutation of three distributions sans negative binomial on *PercLate*. Three permutations of three would normally give us with 27 distinct selections. However, the negative binomial could not be fit on *PercLate* because the former requires the observations (domains) be in whole numbers but the latter contains values that are not, leading to issues with estimating maximum likelihoods of the parameters. Thus we are left with 18 different permutations of distributions.

Another changes made was that since Gamma distribution's domain is $\in (0, \infty)$, under regular circumstances none of our measurements can be fit into any mixture models involving Gamma. Therefore, it was necessary to modify all measurements with values of 0 artificially to an arbitrarily small number, 0.1×10^{-9}, ONLY WHEN a mixture of Gamma distributions was being applied. For instance, in a case where we are fitting **Gamma, Negative Binomial, Gamma** to *Immun*, *NoShow*, and *PercLate*, respectively, then all of *Immun* and *PercLate* whose values were 0 were modified to 0.1×10^{-9}, while *NoShow* was not modified.

Table 3: Log-likelihoods by number of class membership for each permutation of distributions.

Distributions by Variable			**Log-Likelihoods by Number of Class Membership**				
Immun	***NoShow***	***PercLate***	**1**	**2**	**3**	**4**	**5**
Gamma	**Gamma**	**Gamma**	**52776.77**	**56330.92**	**56381.68**	**56707.74**	**57094.26**
Negative Binomial	Gamma	Gamma	45129.16	45830.75	45755.59	46177.64	45993.15
Normal	Gamma	Gamma	44016.88	44695.27	45256.98	45456.71	45704.73
Gamma	Negative Binomial	Gamma	17320.50	21084.25	20779.11	20898.18	20806.93
Negative Binomial	Negative Binomial	Gamma	8411.17	9796.97	10124.40	10252.84	10038.60
Normal	Negative Binomial	Gamma	7298.89	9473.13	9606.42	9868.13	9873.41
gamma	normal	gamma	13902.54	17411.67	19434.70	19334.21	19364.85
negative binomial	normal	gamma	4993.21	9034.48	9132.48	8926.74	9014.72
normal	normal	gamma	2990.93	5201.71	8299.90	8738.49	8445.19
gamma	gamma	normal	5085.02	12387.43	12446.52	14721.77	14016.48
negative binomial	gamma	normal	-3555.20	1863.08	3484.08	11802.22	7422.91
normal	gamma	normal	-4685.09	-3103.42	1668.50	2600.69	3075.54
gamma	negative binomial	normal	-42184.53	-23671.52	-21208.10	-18429.18	-20097.85
negative binomial	negative binomial	normal	-38018.85	-32976.84	-32960.69	-29724.24	-27350.38
normal	negative binomial	normal	-39148.73	-38718.48	-33592.48	-28520.81	-30264.28
gamma	normal	normal	-45603.28	-39752.79	-26035.14	-22972.65	-23436.47
negative binomial	normal	normal	-41437.61	-39687.47	-34566.58	-33855.92	-33784.56
normal	normal	normal	-43456.70	-39362.80	-39136.39	-34824.12	-33398.16

The highest log-likelihoods on each number of components, $K = 1, ..., 5$, was bolded, as well as the permutation that returned the highest overall log-likelihood. For all number of components, the permutation with the highest log-likelihood was **Gamma, Gamma, Gamma**. The highest overall log-likelihood occured at $K = 5$.

We decided to also observe Bayesian Information Criterion for each of them. Figure 3 is the plot of BICs by number of component membership for the five permutations of distributions that were found to have the lowest minimum BICs.

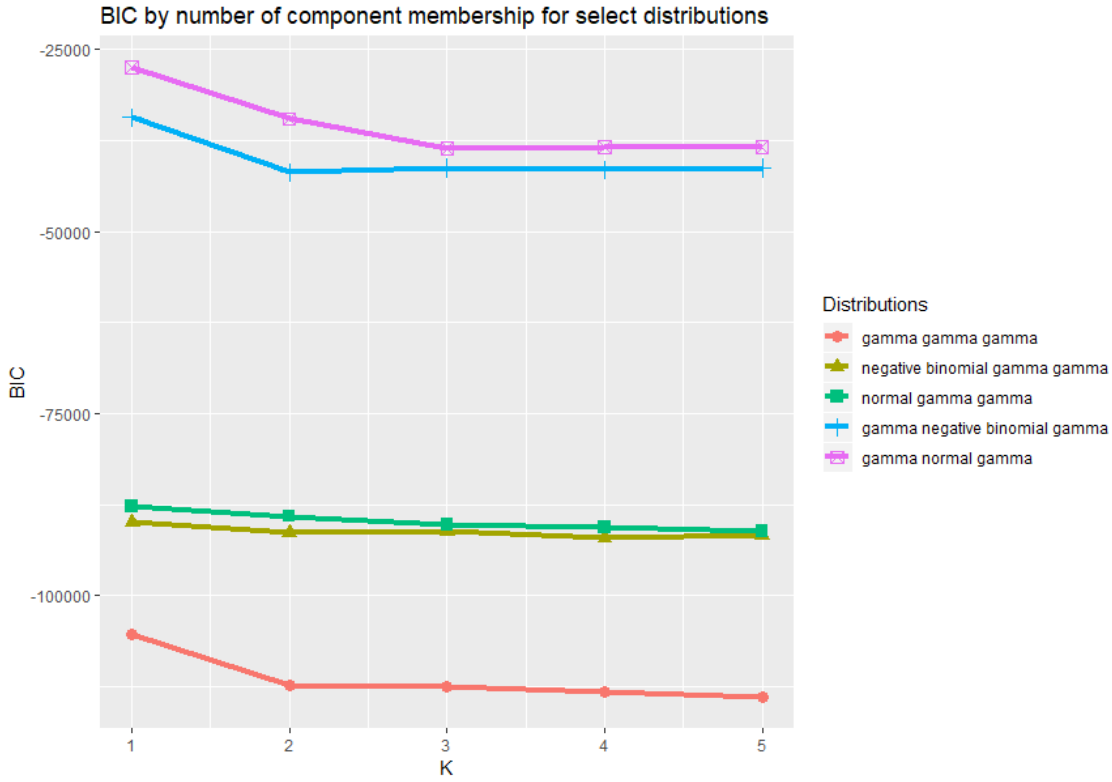

Figure 3: BICs by number of component membership for the five permutations of distributions with the lowest minimum BICs

Figure 4: BICs by number of component membership for the best permutation, **Gamma, Gamma, Gamma**

Not surprisingly, **Gamma, Gamma, Gamma** had the lowest BICs of all permutations at all component memberships. Overall, **Gamma, Gamma, Gamma** at $K = 5$ had the lowest BIC, although $K = 2$, $K = 3$ and $K = 4$ were all nearly as low. One can notice that all five permutations list Gamma as the distribution of choice for *PercLate*, while negative binomial was nearly as often preferred as Gamma for *Immun* and *NoShow*. Gaussian (normal) appears to have been the least frequent choice in any of the measurements - but especially for *PercLate* due to the fact that their density having a domain of $\in (-\infty, \infty)$, although none of the measurements had negative values. Despite this, in some instances, Gaussian distribution was shown to be nearly as competent as the other distributions, as evidenced by the BIC and the log-likelihood and BIC outputs of permutation **Normal, Gamma, Gamma**.

3.3 DENSITY ESTIMATION AND CLUSTERING

In this section we outline key outputs obtained from the EM algorithm, which include maximum likelihood estimates of all relevant parameters, probability

mass and density functions, and the EM clusters derived from the maximum a-posteriori probability estimates. We specifically consider two different models:

- **Gamma, Gamma, Gamma**, $\mathbf{K = 5}$ ($GGGK5$). This is the model with the highest log-likelihoods and the lowest BIC of all the models tested as seen in Table 3 and Figure 3. Sample size is $4,841$, consistent with the number we originally used.
- **Gamma, Gamma, Gamma**, $\mathbf{K = 3}$ ($GGGK3$). This model had the third highest log-likelihoods and the fourth lowest BIC of all the models tested. It was analyzed to see if it successfully overcomes the shortfalls of the first model. Sample size is still $4,841$

Further manipulation to the data set, such as further removing outliers, was initially considered, but was abandoned when it was discovered that it did not improve performance of the EM algorithm.

3.3.1 MODEL 1 - *GGGK5*

The first model to examine is **Gamma, Gamma, Gamma** with five classes (GGGK5), where we obtained the highest log-likelihood and the lowest BIC. This model treated all three measurements as following Gamma distributions with five class memberships.

From our best permutation of distributions and best number of component membership, we obtained the maximum likelihood estimates of all relevant parameters. Estimates of the parameters, $\widehat{\boldsymbol{\alpha}}$ and $\widehat{\boldsymbol{\beta}}$, as well as the proportions of each class membership $\widehat{\pi}_k$, are listed in the Section. From those estimates, we generated a new data set with a dimension of $N \times 3$, where N represents the sample size of $4,841$, identical to the sample size of our data set, and 3 represents the three measurements *Immun*, *NoShow*, and *PercLate*. We call the new data set as the "Expected" values of our observations. Figure 5 and Figure 6 are the histograms and empirical cumulative distribution functions of

the three measurements. The figures will compare the expected and observed values of our measurements and examine performance of our selection.

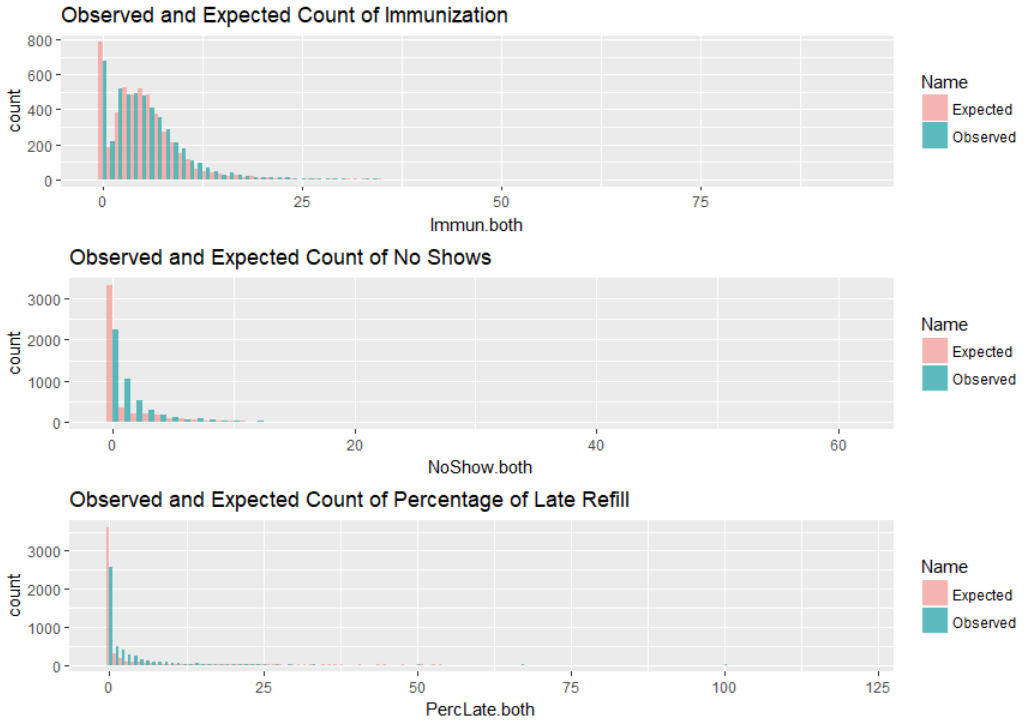

Figure 5: Histograms of expected vs. observed count of the three measurements as estimated by $GGGK5$

Based on the histograms of Figure 5, our model performed the best for *Immun*, where our model did underestimate count at zero, but estimates at other values were very similar to the observed data. Our model performed less well for the other two measurements. For both *NoShow* and *PercLate*, our model grossly overestimated at zero and underestimated elsewhere.

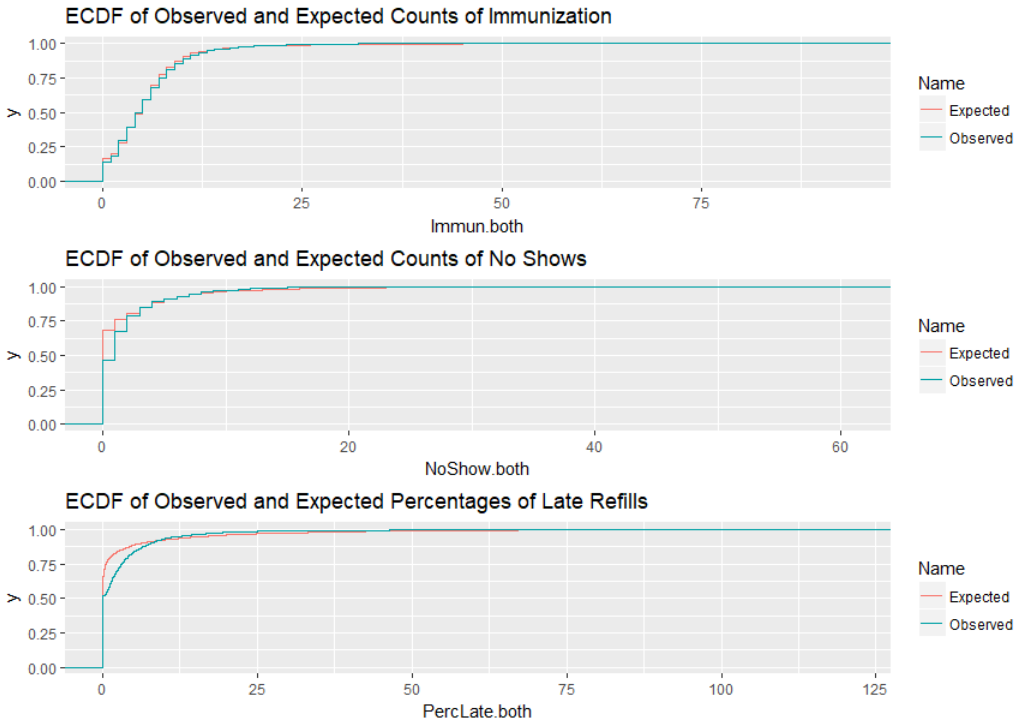

Figure 6: Empirical distribution functions of expected vs. observed count of the three measurements as estimated by $GGGK5$

The empirical distribution functions confirm the results from Figure 5. The model performed reasonably well for *Immun*. However, one can notice that it overestimated at zero and underestimated at most other values for both *NoShow* and *PercLate*.

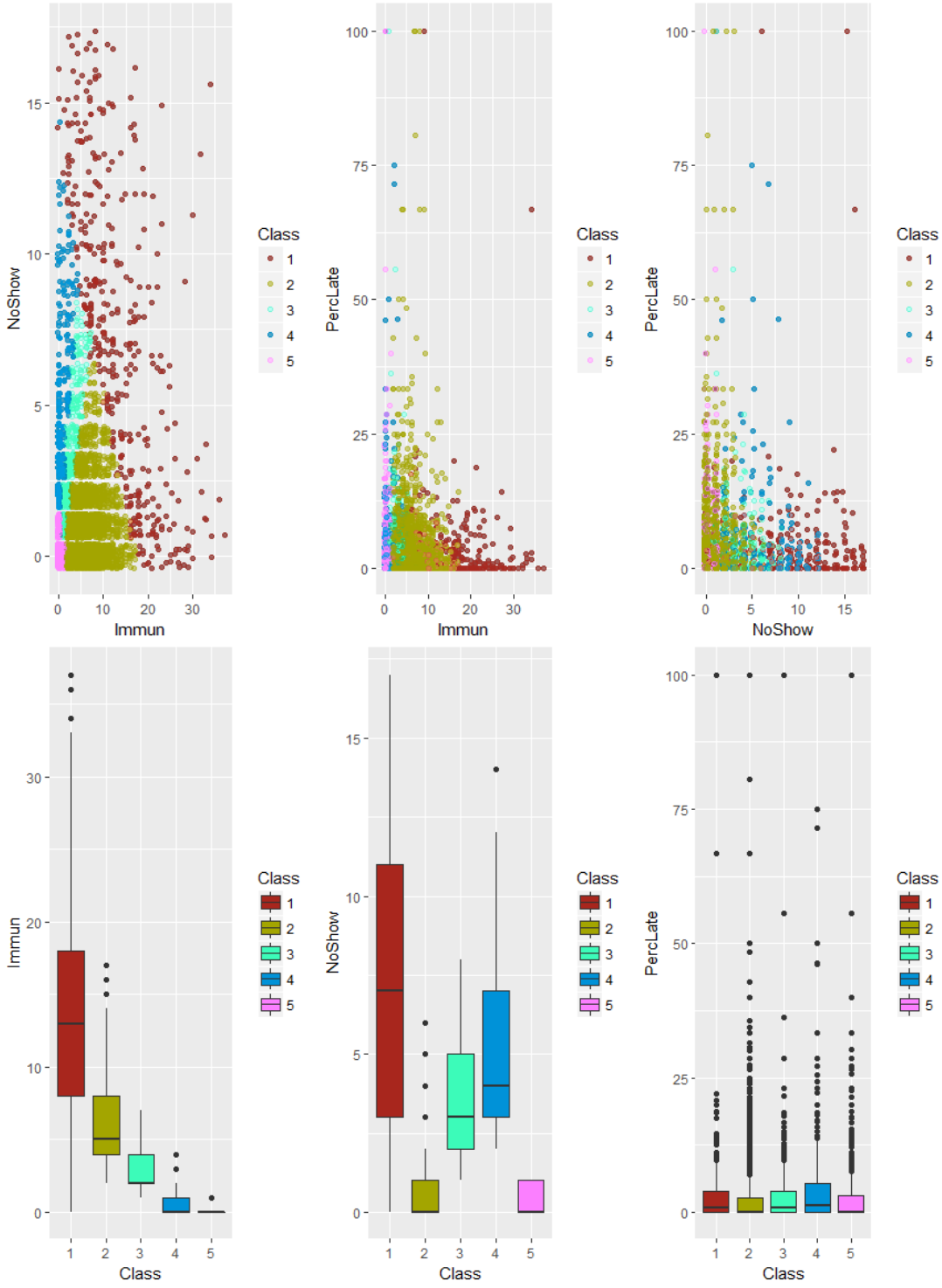

Figure 7: Top: pairwise scatter plots of *Immun*, *NoShow*, and *PercLate* grouped by class
Bottom: boxplots of the aforementioned three measurements grouped by class

Model-based clustering Our model of choice is used to assign observations to clusters that it determines to be most closely associated. Since $K = 5$, the model split our observations into five different clusters. Figures 7 illustrates the pairwise scatter plots and box plots of the three measurements, with the observations grouped by assigned classes $k = 1$, 2, 3, and 4. The scatter plots of discrete measurements were jittered.

Both the scatter plots and the box plots indicate that our model was able to create clear separations among all five classes on *Immun* and *NoShow*. However, separations were close to non-existent on *PercLate*. Based on the clusters generated by this model, we can interpret each cluster as follows:

- Class 1: Patients with very high number of immunizations and no-shows to appointments;
- Class 2: Patients with high number of immunizations and low number of no-shows to appointments;
- Class 3: Patients with medium number of immunizations and no-shows;
- Class 4: Patients with low number of immunizations and high number of no-shows to appointments; and
- Class 5: Patients with very low number of immunizations and no-shows to appointments.

A problem arises when we attempt to rearrange our classes into ordinal variables. Intuitively, a patient with higher engagement levels should have high immunization, low no-shows, and low late medication refill rates, and a patient with low engagement levels should have low immunization, high no-shows, and high late medication refill rates. However, our classes do not align nicely into those characteristics. For instance, *Class 1* has both very high number of immunizations and no-shows to appointments, whereas *Class 5* has both very low number of immunizations and no-shows to appointments, generating a rather contradictory outcomes.

3.3.2 MODEL 2 - *GGGK3*

Due to a shortcoming of the model with five classes, we examined the model with three classes $GGGK3$. Because there are fewer classes, we hoped that the clusters would follow a more ordinal fashion.

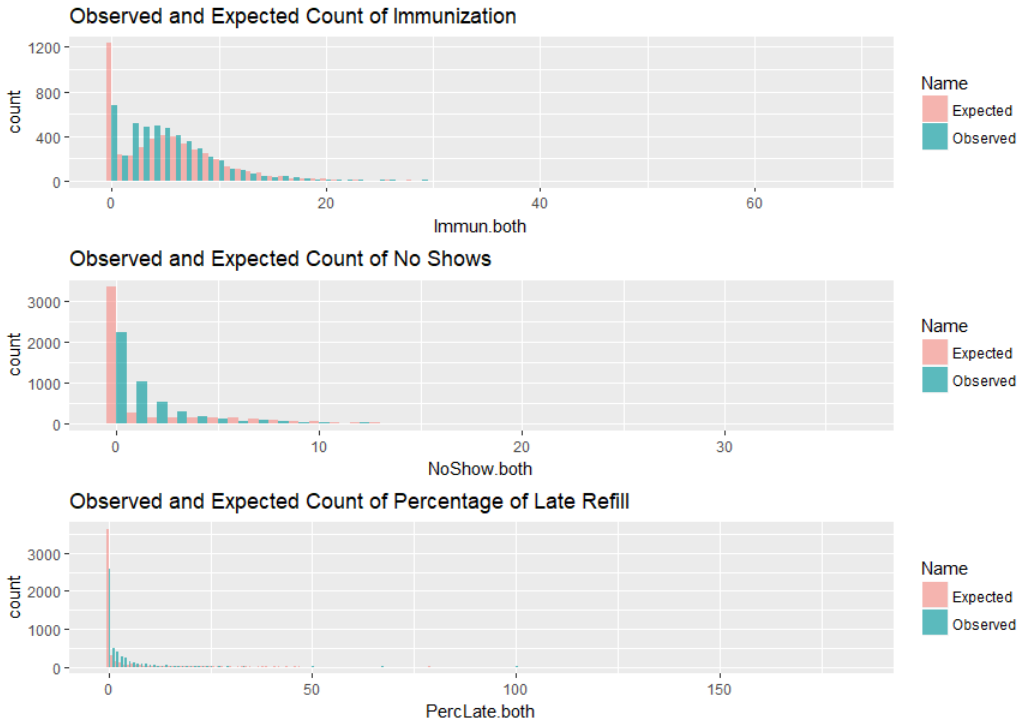

Figure 8: Histogram of expected vs. observed count of the three measurements as estimated by the model $GGGK3$

Figure 9: eCDF of expected vs. observed count of the three measurements as estimated by the model $GGGK3$

Based on Figure 8 and Figure 9, the model $GGGK3$ performed somewhat worse than our first model, $GGGK5$. This model overestimated the count of *Immun* at the value of 0 at a far higher frequency while underestimating at everywhere else with even greater degree than the previous model. For *NoShow* and *PercLate*, the model also overestimated at 0 and underestimated at other values but at a greater magnitude.

Model-based clustering Both the pairwise scatter plots and the box plots in Figure 10 demonstrate that *PercLate* played little role in determining class memberships of the observations, mirroring the conclusion we reached for the earlier model. Similarly to the previous model, our new model created some visible intra-class separations on *Immun* and *NoShow*, but at a smaller magnitude. For instance, instead of clear five separate classes, now we have only one class - *Class 1* on *Immun* and *Class 3* on *NoShow* - being visibly distinct from

the others.

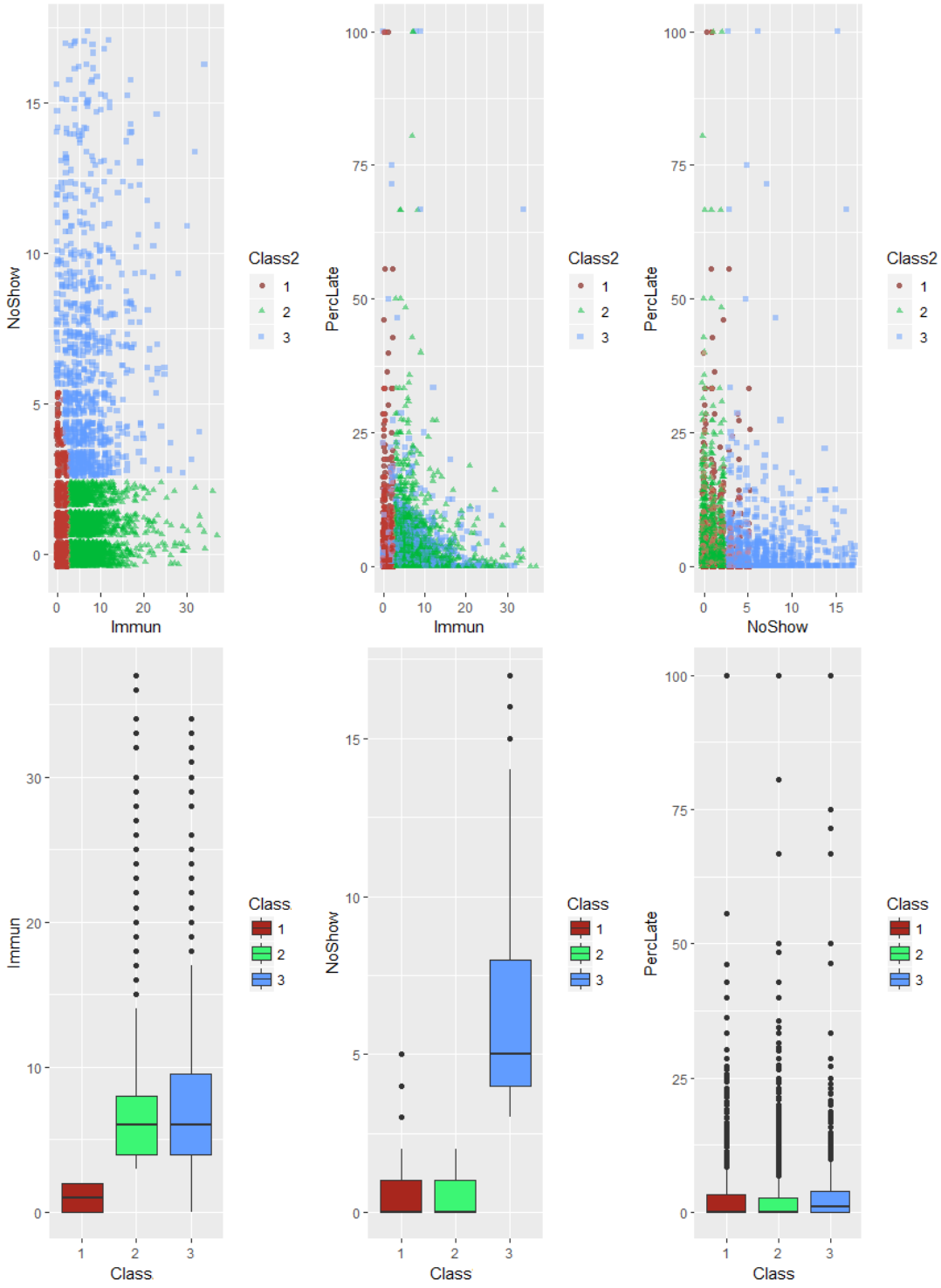

Figure 10: Top: pairwise scatter plots of *Immun*, *NoShow*, and *PercLate* grouped by class
Bottom: box plots of the three measurements grouped by class for $GGGK3$ model

Table 4: Summary output of simple linear regression of the measurements vs. class membership.[5]

Immun

	Class 1	*Class 3*
Class 2	✓	✗
Class 3	✓	-

NoShow

	Class 1	*Class 3*
Class 2	✓	✓
Class 3	✓	-

PercLate

	Class 1	*Class 3*
Class 2	✓	✓
Class 3	✓	-

To test our findings from above further, we performed simple linear regression among all the classes versus each of the three measurements. Our null hypothesis is that the mean measurements of all three classes were the same for all three measurements. The output can be seen in Section A, while Table 4 shows the simpler version of the summary output. Contrary to our visual analysis, all but one comparison - between $k = 1$ and $k = 3$ of *Immun* - had p-value less than 0.05. Most notably, there were evidence that there are differences in mean values among clusters on *PercLate*. *Class 3* had the largest mean percentage of late medication refills while *Class 2* had the smallest mean percentage of late medication refills. Despite this, there were huge overlap, especially with outliers, among all classes on *PercLate*. Also, this regression assumes that the response measurements follow normal distributions, which we already know to be not true, though this concern is somewhat muted due to large sample size rendering assumption of normality of residuals less of a concern.[4] As a result, we tried negative inverse regression by assuming the response measurements each follow a single Gamma distribution.

[5] Classes where differences of means were shown to be significant (p-value < 0.05) are marked as ✓.

Table 5: Summary output of generalized linear regression (negative inverse link function) of the measurements vs. class membership.[6]

Immun

	Class 1	*Class 3*
Class 2	✓	✗
Class 3	✓	-

NoShow

	Class 1	*Class 3*
Class 2	✓	✓
Class 3	✓	-

PercLate

	Class 1	*Class 3*
Class 2	✓	✓
Class 3	✓	-

Table 5 demonstrates that conclusions we reached for the $\hat{\beta}$ parameters have not changed even after the application of negative inverse link function on the generalized linear regression. However, this model also has its shortcomings as we are already aware that the measurements do not follow a single Gamma distribution. Thus one should proceed with caution when inferring meaningful conclusion from the summary outputs of both regression models. The models do not imply EM algorithm's effectiveness.

Based on the clusters generated by this model, we can interpret each cluster as follows:

- Class 1: Patients with low number of immunizations no-shows to appointments;
- Class 2: Patients with high number of immunizations and low number of no-shows to appointments; and
- Class 3: Patients with high number of immunizations and no-shows to appointments.

The classes are then sorted by the measurement that had the highest weighted intra-class variance. In this case, that measurement was *NoShow*, with the weighted intra-class variance of 0.6253. Figure 11 shows the box plots of the

[6] Classes where differences of means were shown to be significant (p-value < 0.05) are marked as ✓.

measurements with classes reordered by *NoShow* in descending order because higher values in *NoShow* intuitively represent lower patient engagement level.

Table 6: Measurements by weighted intra-class, inter-class, and total variances

	Immun	***NoShow***	***PercLate***
$\boldsymbol{\sigma}_d^2(intraclass)$	0.3316	**0.6253**	0.0034
$\boldsymbol{\sigma}_d^2(interclass)$	0.6684	0.3747	0.9966
$\boldsymbol{\sigma}_d^2(total)$	1.0000	1.0000	1.0000

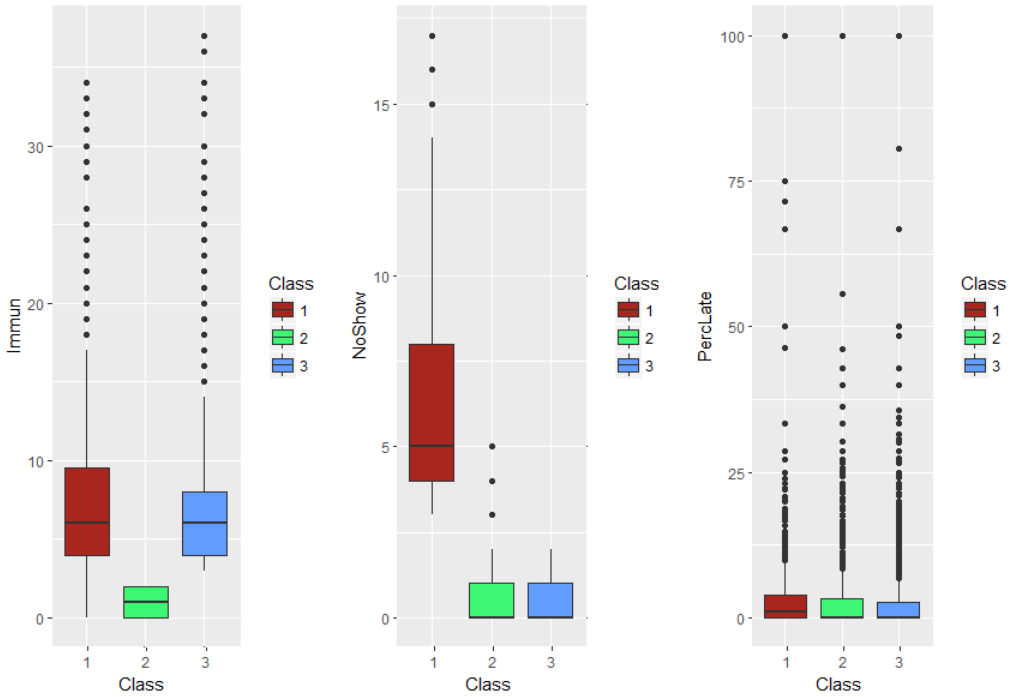

Figure 11: Box plots of the three measurements grouped by class[7]

Conveniently, reordering classes by *NoShow* in descending order also sorts them by *PercLate* in the same order. The same could not be said of *Immun*, where *Class 3* has higher mean *Immun* values than the *Class 2* does. The new classes will now be referred to as **Patient Engagement Score (PES)**, where classes 1, 2 and 3 represent patients with the lowest, medium, and the highest patient engagement level, respectively.

[7]Classes were reordered by *NoShow* in descending order

We performed multiple logistic regression for the association between this PES and health-related outcome measurements adjusted by patient characteristics and health behavior.

3.3.3 SUMMARY

In this section we analyzed our raw data, measurements used, and the types of models developed. Though a total of 36 models were developed through the application of EM algorithm - 18 different permutations of distribution types with outliers and 18 without - we specifically delved into two different models:

1. **Gamma, Gamma, Gamma,** $K = 5$ - $GGGK5$;
2. **Gamma, Gamma, Gamma,** $K = 3$ - $GGGK3$

These models were chosen due to low BIC levels and/or relative simplicity in interpretation. The first model, **Gamma, Gamma, Gamma,** $K = 5$ -$GGGK5$ was not analyzed further due to difficulty in arranging the class membership into ordinal fashion. Outputs of the second model will be discussed in the next subsection, Section 3.4.

3.4 MULTIPLE LOGISTIC REGRESSION ANALYSIS

In this section, we will discuss outputs from the multiple logistic regression for the association between two different Patient Engagement Scores generated from one of the two models developed respectively and health-related outcome measurements adjusted by patient characteristics and health behavior. The models used was **Gamma, Gamma, Gamma**, $\mathbf{K = 3}$ ($GGGK3$).

Table 7: Summary of health-related outcome measurements

Variables	**"0"**		**"1"**	
HDL	3,174	(normal)	1,667	(dyslipidemia)
LDL	2,924	(normal)	1,917	(dyslipidemia)
EDVisit	3,152	(none)	1,689	(ED visited)
A1C	3,174	(normal)	1,667	(diabetes)
eFGR	1,337	(normal)	3,504	(not normal)
Hospitalization	3,595	(none)	1,246	(hospitalized)
SBP	4,313	(normal)	528	(hypertension)
DBP	4,618	(normal)	223	(hypertension)

Table 8: Summary of eight measurements on patient characteristics and health-related behavior (control variables)

Control Variables	**"0"**		**"1"**	
Tobacco	626	(smoker)	4,215	(non-smoker)
Alcohol	2,741	(yes)	2,100	(no)
Marital	1,386	(not married)	3,455	(married)
Race	4,595	(white)	246	(non-white)
Gender	2,616	(male)	2,225	(female)
Primary Insurance	4,057	(non-govt)	784	(govt)

Control Variable	**Min.**	**Median**	**Max.**	**Mean**	**SD**
Age	18	59	96	57.24	10.03
Chronic Conditions	2	3	21	3.239	1.670

Table 7 is a summary of health-related outcome measurements, the dependent variables in multiple logistic regression, whereas Table 8 summarizes the eight measurements on patient characteristics and health-related behavior, which serve as the control variables. All eight health-related outcome measurements are binary values, where zero indicates absence of a given medical condition and one indicates presence. As such, zero is preferable from a patient's perspective. Out of the eight control variables, six - *Tobacco*, *Alcohol*, *Marital*, *Race*, *Gender*, and *Primary Insurance* - are binaries and two - *Age* and *Chronic Condition* are strictly natural numbers. In contrast to the health-related outcome variables, zeros and ones do not necessarily correspond to desirability in control variables.

	PAM			
PES	1	2	3	4
1	3	3	5	11
2	1	2	1	3
3	2	9	8	17

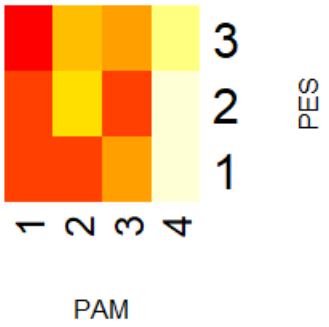

Figure 12: PES vs. PAM by counts for $GGGK3$ results [8]

PES validation One of the ways the new score can be validated it to see if it has association with the PAM score. To assess this, a contingency table and heatmap is constructed for those patients who had both PAM and PES score. The result can be found in Figure 12, In total there were only 65 patients that had both PES and PAM scores. This sample size is too small to make any meaningful inference based on Pearson's chi-squared test; however the current results sugesst that PES and PAM were not very closely related. Chi-squared test was not performed due to limited sample size of patients with PAM assigned and would have failed to produced reliable output. This result can due sample size. In addition, we also noted that PAM was not significantly associated with the initial three indirect measures used to construct the new score.

[8] The heatmap's intensity (brightness) represents proportion of each square by PES

Table 9: Patient characteristic and health behavior measurements that were found to have significant association by health-related outcomes; numbers in parentheses represent odds ratio

	SBP	**DBP**	**HDL**	**LDL**
P-value $<$ **0.05**	Age (1.021) BMI (1.026) Tobacco (0.597)	Age (0.969) Gender (0.532) BMI (1.038)	**PES3 (0.814)** Age (0.985) Gender (0.291) BMI (1.040) Tobacco (0.582) Alcohol (1.444) ChrCond (1.125)	**PES3 (0.756)** Age (0.983) Gender (1.627) ChrCond (0.860)
	ED Visit	**A1C Level**	**eGFR**	**Hospitalization**
P-value $<$ **0.05**	**PES1 (3.071)** Alcohol (1.256) ChrCond (1.166) Age (0.990)	Age (0.859) Alcohol (1.360) ChrCond (1.249) BMI (1.012)	Age (1.066) Gender (1.790) Tobacco (1.582) BMI (0.986)	**PES1 (3.508)** Gender (0.829) Alcohol (1.256) PrimInsur (1.344) ChrCond (1.201)

The second method we can use to validate the new score is to analyze it association with health-related outcomes. Table 9 shows the key outputs from the multiple logistic regression. Note that PES 2 was used as the reference level. For instance, *Age* was found to have an odds ratio of 1.021 on *Systolic BP*. Interpretation can be given that, keeping any other predictors constant, one year increase in patient's age was associated with 1.021 times (alternatively, 2.1 % increase in) the probability of the patient suffering from hypertension. Details of the results is reported in the Appendix (Section A).

The multiple logistic regression was able to detect potential association between the PES we created and some health-related outcome measurements. Patients who were assigned PES score of 3 (highest level of engagement) were significantly less likely to be suffering from both types of dyslipidemia (HDL and LDL) than those who were assigned different PES. Likewise, patients who were assigned PES score of 1 (lowest level of engagement) were significantly more likely to have experienced emergency department visit and hospitalization than those of other PES groups. None of these conclusions were contradictory to our expectations.

Table 10: Comparison of output between Ngorsuraches et al.'s model and *GGGK*3

	Ngorsuraches et al. (Index = PES 3)	GGGK3 (Index = PES 2)
Significant correct prediction	LDL (PES 1) ED Visit (PES 2) A1C (PES2) eGFR (PES 1) Hospitalization (PES 2)	HDL (PES 3) LDL (PES 3) ED Visit (PES 1) Hospitalization (PES 1)
Significant conflicting prediction	LDL (PES 2)	

Recall that in Section 1, the purpose of the project was to improve on the model developed by Ngorsuraches et al. Table 10 summarizes significant predictions obtained by Ngorsuraches' model and *GGGK*3. Ngorsuraches et al. had also assigned each patient with one of three preliminary scores, also referred to as PES, with higher score representing higher level of patient engagement. Ngorsuraches' model found that patients assigned to PES 1 were associated with higher risk of experiencing dyslipidemia per LDL and abnormal renal function per eFGR, whereas patients assigned to PES 2 were associated with higher risk of having experienced ED visit and hospitalization, as well as higher odds of suffering from diabetes per A1C level than those in PES 3. Overall, Ngorsuraches' model was able to correctly predict more health outcomes than *GGGK*3 did.

However, Ngorsuraches' model also found that patients in PES 2 had lower risk of experiencing dyslipidemia per LDL than those in PES 3, which is contradictory to our expectation. Further, ED visit, A1C, and hospitalization were found to be significantly higher only for PES 2 compared to PES 3 but not for PES 1, suggesting further contradiction. In contrast, no such contradiction was found in *GGGK*3.

3.4.1 SUMMARY

In this section, we discussed the outputs of multiple logistic regression for the association between PES generated by **Gamma, Gamma, Gamma, K = 3** - $GGGK3$ model developed in the earlier section. This model was based on a sample size of 4,841 patient three indirect measures of engagement. Eight different regressions were performed, where each of them represents one of the eight health-related outcome measurements - SystolicBP, DiastolicBP, HDL, LDL, EDVisit, A1C, eFGR, and Hospitalization. All of the eight were converted into binary outputs, where zero generally represents the default, or the preferable outcome, whereas one generally represents presence of issues.

This model suggests that, keeping control variables constant, patients assigned to PES of 3 had lower odds of experiencing dyslipidemia in terms of both HDL and LDL than those assigned different PES, and the patients assigned 1 had elevated odds of having experienced emergency department visits and hospitalization lasting more than one day than those with different PES. As higher PES are expected to represent improved patient engagement level, these outcomes are promising.

4 DISCUSSION

There has been numerous research that corroborate link patient engagement levels with improved health outcomes. However, existing measures that estimate the patient engagement levels rely on subjective inputs of patients and health care professionals. In this paper, we described data-driven patient engagement score. This is an instrument that could be used for calculating patient engagement levels strictly from a small number of easily-obtainable, objective health and behavior related measures of patients.

Three indirect measurements of engagement were selected - number of immunizations received, no shows on appointments with medical personnel, and

percent of medication refills achieved after deadline as indirect measures of patient engagement. Finite mixture modeling technique was applied to the three measurements to generate a new latent variable, which was aimed to represent patient engagement levels. We call the new score - Patient Engagement Scores or PES.

In Section 3, we explored the data set and selected the best fitting models. Two competing models were considered: Model 1. five-component mixture model, which we refer to as $GGGK5$; and Model 2. three-component mixture model, referred to as $GGGK3$. Finally, the best model was determined based on two criterion; 1. interpretability - whether the PES generated were in ordinal fashion and give meaningful groups - and 2. predictability - whether the PES showed association with outcome variables based on the multiple logistic regression.

Based on the interpretability, we determined that the Model 1, $GGGK5$, did not meet our criteria as practical. This is because it contained five distinct classes, none of which were easily reorganized as an ordinal variable. Model 2, $GGGK3$, created reasonable ordinal alignment for at least two of the three measurements and were able to generate a more interpretable outputs. Therefore, only Model 2, $GGGK3$, was considered for the regression. According to this model, after adjusting for the patient characteristic and health behavior metrics, higher PES was found to be associated with reduced odds of dyslipidemia, and lower PES was associated with elevated odds of experiencing emergency department visit and hospitalization within the last three years.

Comparison with the model created in a previous study by Ngorsuraches et al. returned mixed result. Ngorsuraches' model had been generated through application of multivariate Gaussian mixture model. Preliminary scores generated by Ngorsuraches' model was able to correctly predict more health-related outcomes than $GGGK3$, but also produced conflicting predictions. In contrast, $GGGK3$ did not produce any contradictory result.

Our method of predicting patient engagement has several advantages. In contrast to PAM, which relies heavily on subjective outputs from the patients and the medical professional, PES relies almost exclusively to a select few objective measurable data. Due to time-consuming and often expensive nature of the instrument, PAM also suffers from reduced applicability. In our data set with a total sample size of $147,687$ but only $1,442$ observations had available PAM scores. In contrast, $103,686$ of the observations contained information regarding the three measurements, indicating that most observations in the data set are eligible for a PES.

However, there are many caveats and potential issues with our methodology. Firstly, outliers from each measurement had already been removed from the data set. It is currently unknown if our model would have yielded a very different outcomes had those outliers have been included. In addition, in both models, only a small subset of the observations were selected. This was necessary as we needed complete information for our regression models. This could become an issue if patients with incomplete information were found to be significantly different in some patient characteristics or health behavior measurements than those with complete information.

Another limitation is the model's initial assumptions. When fitting a mixture model, it is assumed that the variables are independent of one another. However, initial investigation suggested this assumption is not guaranteed. This finding is not surprising - or arguably even desirable - if they indeed were associated with patient engagement levels.

4.1 FUTURE STUDIES

While the outcomes of PES were promising, there were some items that could be topics of future studies. As mentioned in Section 1, a study by Rahimi et al. reported that current studies with the objective of developing a tool measuring patient engagement suffer from lack of measurement properties, which include

structural and criterion validity, consistency, reliability, and responsiveness.[18] Our study was, likewise, unable to satisfy some of these measurement properties due to lack of necessary resources.

A competent measurement should be able to predict patient engagement levels of the patients from other geographical regions and with different health care providers. This could be further work using other dataset to validate this as our patient data set was obtained exclusively from Sanford Health, which operates primarily in the West North Central region of the United States. Further, other patient behavioral measures that are reflective of engagement such as patient initiated visits, length of visit, patient portal use through log files could be considered to make the score more robust.

A future study would therefore involve validating the model's measurement properties, testing the model on a different set of patients to confirm if it still meets the criteria for those with significantly different backgrounds. In addition, responsiveness can be incorporated into the model to improve detection of changes occurring to patient engagement levels, potentially through the use of time series analysis. Finally, additional data regarding patient portal uses and other variables will help improve the current model. This can be achieved as more health care providers adopt inpatient and outpatient portal, increasing amount of data regarding portal user behaviors of patients become more developed.

A APPENDIX

A.1 THE EM ALGORITHM FOR THE MIXTURE MODELS

In Section 2, The finite mixture model is given as if independence is assumed between the multi-variable data:

$$g(\boldsymbol{y};\boldsymbol{\psi}) = \sum_{k=1}^{K} \pi_k \prod_{d=1}^{D} f_d(\boldsymbol{y_{id}};\boldsymbol{\theta}_{kd}).$$

Here we have three variables under consideration hence mixture distributions can be written as follows

$$g(\boldsymbol{y}_i;\boldsymbol{\psi}) = \sum_{k=1}^{K} \pi_k \prod_{d=1}^{3} f_d(\boldsymbol{y_{id}};\boldsymbol{\theta}_{kd}).$$

If we consider the three measurements followed a mixture of distributions of **Gaussian**, **Negative Binomial**, and **Gamma**, in that order, the mixture will have the following form

$$g(\boldsymbol{y}_i;\boldsymbol{\psi}) = \sum_{k=1}^{K} \pi_k (f_{Gaus}(y_{i1};\mu_k,\sigma_k^2) f_{NB}(y_{i2};r_k,p_k) f_{Gam}(y_{i3};\alpha_k,\beta_k)).$$

Then the log-likelihood function is given by:

$$\log L(\boldsymbol{\psi}|\boldsymbol{y}_i) = \sum_{i=1}^{N} \log(\sum_{k=1}^{K} \pi_k (f_{Gaus}(y_{i1};\mu_k,\sigma_k^2) f_{NB}(y_{i2};r_k,p_k) f_{Gam}(y_{i3};\alpha_k,\beta_k))).$$

Suppose we know z_i indicating the true component membership of th i^{th} observations. Then the complete data likelihood is

$$L_c(\boldsymbol{\psi}|\boldsymbol{y}_i) = \prod_{i=1}^{N} \prod_{k=1}^{K} (\pi_k \prod_{d=1}^{D} f_d(\boldsymbol{y}_{id};\boldsymbol{\theta}_{kd}))^{I(z_i=k)}.$$

The complete data log likelihood becomes

$$\log L_c(\boldsymbol{\psi}|\boldsymbol{y}_i) = \ell_c(\boldsymbol{\psi}|\boldsymbol{y}_i) = \sum_{i=1}^{N}\sum_{k=1}^{K} I(z_i = k) \sum_{d=1}^{D} \log(\pi_k f_d(\boldsymbol{y}_{id}; \boldsymbol{\theta}_{kd})).$$

In the EM algorithm we first do the conditional expectation of the complete data log-likelihood function which gives the function as

$$\begin{aligned}
Q(\boldsymbol{\psi}|\boldsymbol{\psi}^{(s)}) &= \mathbb{E}_{\boldsymbol{Z}|\boldsymbol{Y},\boldsymbol{\psi}^{(s)}}[\log L_c(\boldsymbol{\psi}|\boldsymbol{Y}, \boldsymbol{Z})] \\
&= \mathbb{E}_{\boldsymbol{Z}|\boldsymbol{Y},\boldsymbol{\psi}^{(s)}} \left[\sum_{i=1}^{N} \log L_c(\boldsymbol{\psi}|\boldsymbol{y}_i, \boldsymbol{z}_i)\right] \\
&= \sum_{i=1}^{N} \mathbb{E}_{\boldsymbol{Z}|\boldsymbol{Y},\boldsymbol{\psi}^{(s)}}[\log L_c(\boldsymbol{\psi}|\boldsymbol{y}_i, \boldsymbol{z}_i)] \\
&= \sum_{i=1}^{N}\sum_{k=1}^{K} p(z_i = k|\boldsymbol{y}_i; \boldsymbol{\psi}) \log L(\boldsymbol{\psi}_k|\boldsymbol{y}_i, \boldsymbol{z}_i) \\
&= \sum_{i=1}^{N}\sum_{k=1}^{K} p(z_i = k|\boldsymbol{y}_i; \boldsymbol{\psi}) \sum_{d=1}^{D} \log(\pi_k^{(s)} f_d(\boldsymbol{y}_{id}; \boldsymbol{\theta}_{kd})),
\end{aligned}$$

which reduces to find the posterior probability of the component membership. At **E-step** the estimated a-posteriori probabilities follows the following form:

$$\begin{aligned}
\tau_{ik}^{(s+1)} &= p(z_i = k|\boldsymbol{y}_i; \boldsymbol{\psi}) \\
&= \frac{\pi_k^{(s)} p(\boldsymbol{y}_i|z_i = k; \boldsymbol{\theta}_k)}{\sum_{k=1}^{K} \pi_k^{(s)} p(\boldsymbol{y}_i|z_i = k; \boldsymbol{\theta}_k)} \\
&= \frac{\pi_k^{(s)} f(\boldsymbol{y_i}; \boldsymbol{\theta}_k^{(s)})}{\sum_{k=1}^{K} \pi_k^{(s)} f(\boldsymbol{y_i}; \boldsymbol{\theta}_k^{(s)})} \\
&= \frac{\pi_k^{(s)} \prod_{d=1}^{D} f_d(\boldsymbol{y}_{id}; \boldsymbol{\theta}_{kd}^{(s)})}{\sum_{k=1}^{K} \pi_k^{(s)} \prod_{d=1}^{D} f_d(\boldsymbol{y}_{id}; \boldsymbol{\theta}_{kd}^{(s)})}.
\end{aligned}$$

The Q-function becomes

$$\begin{aligned}
Q(\boldsymbol{\psi}^{(s+1)}|\boldsymbol{\psi}^{(s)}) &= \sum_{i=1}^{N}\sum_{k=1}^{K}\widehat{\tau}_{ik}^{(s+1)}\sum_{d=1}^{D}\log(\widehat{\pi}_k^{(s+1)}f_d(\boldsymbol{y}_{id};\boldsymbol{\theta}_{kd}^{(s+1)})) \\
&= \sum_{i=1}^{N}\sum_{k=1}^{K}\widehat{\tau}_{ik}^{(s+1)}\left(\sum_{d=1}^{D}\log(f_d(\boldsymbol{y}_{id};\boldsymbol{\theta}_{kd}^{(s+1)})) + \log\widehat{\pi}_k^{(s+1)}\right) \\
&= \sum_{i=1}^{N}\sum_{k=1}^{K}\widehat{\tau}_{ik}^{(s+1)}\sum_{d=1}^{D}\log(f_d(\boldsymbol{y}_{id};\boldsymbol{\theta}_{kd}^{(s+1)})) + \sum_{i=1}^{N}\sum_{k=1}^{K}\widehat{\tau}_{ik}^{(s+1)}\log\widehat{\pi}_k^{(s+1)} \\
&= Q_1(\boldsymbol{\theta}^{(s+1)}|\boldsymbol{\psi}^{(s)}) + Q_2(\boldsymbol{\phi}^{(s+1)}|\boldsymbol{\psi}^{(s)}),
\end{aligned}$$

where

$$\begin{aligned}
Q_1(\boldsymbol{\theta}^{(s+1)}|\boldsymbol{\psi}^{(s)}) &= \sum_{i=1}^{N}\sum_{k=1}^{K}\widehat{\tau}_{ik}^{(s+1)}\sum_{d=1}^{D}\log(f_d(\boldsymbol{y}_{id};\boldsymbol{\theta}_{kd}^{(s+1)})) \\
&= \sum_{i=1}^{N}\sum_{k=1}^{K}\widehat{\tau}_{ik}^{(s+1)}\Big[\log(f_{Gaus}(y_{i1};\mu_{k1}^{(s+1)}\sigma_{k1}^{2(s+1)})) \\
&\quad + \log(f_{NB}(y_{i2};r_{k2}^{(s+1)}p_{k2}^{(s+1)})) \\
&\quad + \log(f_{Gam}(y_{i3};\alpha_{k3}^{(s+1)}\beta_{k3}^{(s+1)}))\Big] \\
&= \sum_{i=1}^{N}\sum_{k=1}^{K}\widehat{\tau}_{ik}^{(s+1)}\left[\log\left(\frac{1}{\sigma\sqrt{2\pi}}e^{-(y_{i1}-\mu_k)^2/2\sigma_k^2}\right)\right. \\
&\quad + \log\left(\frac{\Gamma(y_{i2}-r_k)}{\Gamma(r_k)\Gamma(y_{i2}+1)}p_k^{r_2}(1-p_k)^{y_{i2}}\right) \\
&\quad \left. + \log\left(\frac{{y_{i3}}^{\alpha_k-1}e^{-(y_{i3}/\beta_k)}}{{\beta_k}^{\alpha_k}\Gamma(\alpha_k)}\right)\right]; \\
Q_2(\boldsymbol{\phi}^{(s+1)}|\boldsymbol{\psi}^{(s)}) &= \sum_{i=1}^{N}\sum_{k=1}^{K}\widehat{\tau}_{ik}^{(s+1)}\log(\pi_k^{(s+1)}).
\end{aligned}$$

At **M-step** we take partial derivative of the Q_1 and Q_2 -function for each $\boldsymbol{\theta}_{kd}$. In our model, by taking the partial derivatives we find the estimate of

μ_k, σ_k^2

$$\begin{aligned} \frac{\partial Q_1}{\partial \mu_k} &= \frac{\sum_{i=1}^{N} \tau_{ik}(y_{i1} - \mu_k)}{\sigma_k^2} = 0 \\ \implies \sum_{i=1}^{N} \tau_{ik} y_{i1} - \sum_{i=1}^{N} \tau_{ik}\mu_k &= 0 \end{aligned} \tag{A.1}$$

$$\begin{aligned} \frac{\partial Q_1}{\partial \sigma_k^2} &= \sum_{i=1}^{n} \tau_{ik}\left(\frac{-1}{\sigma_k} + \frac{(y_{i1} - \mu_k)^2}{\sigma_k^3}\right) = 0 \\ \implies \frac{\sum_{i=1}^{n} \tau_{ik}}{\sigma_k} &= \frac{\sum_{i=1}^{n} \tau_{ik}(y_{i1} - \mu_k)^2}{\sigma_k^3} \end{aligned} \tag{A.2}$$

Therefore, at the (s+1)th iteration of the M-step of the EM algorithm using Equations A.1 and A.2 we get the following

$$\mu_k^{(s+1)} = \frac{\sum_{i=1}^{N} \tau_{ik}^{(s+1)} y_{i1}}{\sum_{i=1}^{N} \tau_{ik}^{(s+1)}} \text{ and } \sigma_k^{2(s+1)} = \frac{\sum_{i=1}^{n} \tau_{ik}^{(s+1)} (y_{i1} - \mu_k^{(s+1)})^2}{\sum_{i=1}^{n} \tau_{ik}^{(s+1)}}.$$

Similarly, for r_k, p_k, α_k, and β_k, we use the following derivations to find estimates at the (s+1)th iteration of the M-step

$$\begin{aligned} \frac{\partial Q_1}{\partial p_k} &= \frac{\sum_{i=1}^{N} \tau_{ik} y_{i2}}{1 - p_k} - \frac{\sum_{i=1}^{N} \tau_{ik} r_k}{p_k} = 0 \\ \implies \frac{\sum_{i=1}^{N} \tau_{ik} y_{i2}}{1 - p_k} &= \frac{\sum_{i=1}^{N} \tau_{ik} r_k}{p_k} \\ \implies p_k \sum_{i=1}^{N} \tau_{ik} y_{i2} &= \sum_{i=1}^{N} \tau_{ik} r_k - p_k \sum_{i=1}^{N} \tau_{ik} r_k \\ \implies p_k \left(\sum_{i=1}^{N} \tau_{ik} y_{i2} + \sum_{i=1}^{N} \tau_{ik} r_k \right) &= \sum_{i=1}^{N} \tau_{ik} r_k. \end{aligned} \tag{A.3}$$

Therefore, the closed form solution at the (s+1)th iteration is given as

$$p_k^{(s+1)} = \frac{\sum_{i=1}^{N} \tau_{ik}^{(s+1)} r_k^{(s)}}{\sum_{i=1}^{N} \tau_{ik}^{(s+1)} y_{i2} + \sum_{i=1}^{N} \tau_{ik}^{(s+1)} r_k^{(s)}}.$$

For the estimation of $r_k^{(s+1)}$ consider the following

$$\frac{\partial Q_1}{\partial r_k} = \left(\sum_{i=1}^{N} \tau_{ik} r_k \psi(y_{i2} + r_k) \right) + \sum_{i=1}^{N} \tau_{ik} \psi(r_k) + \sum_{i=1}^{N} \tau_{ik} \log(1 - p_k) = 0, \tag{A.4}$$

where $\psi(y_{i2}) = \frac{\Gamma^{'}(y_{i2})}{\Gamma(y_{i2})}$. Using the solution for $p_k^{(s+1)}$ we obtain the following expression

$$\begin{aligned} \frac{\partial Q_1}{\partial r_k} &= \left(\sum_{i=1}^{N} \tau_{ik} r_k \psi(y_{i2} + r_k) \right) + \sum_{i=1}^{N} \tau_{ik} \psi(r_k) \\ &+ \sum_{i=1}^{N} \tau_{ik} \log \left(1 - \frac{\sum_{i=1}^{N} \tau_{ik} r_k}{\sum_{i=1}^{N} \tau_{ik} y_{i2} + \sum_{i=1}^{N} \tau_{ik} r_k} \right) = 0. \end{aligned} \tag{A.5}$$

Equation A.5 cannot be solved in closed form; therefore, at the M-step, the $r_k^{(s+1)}$ cannot be solved in close form. To obtain the parameter estimates numerical approximation method should be applied. We relied on Nelder-Mead method to find the estimate $r_k^{(s+1)}$.

Finally, estimating the parameters of the Gamma distribution can be done applying the following derivations:

$$\frac{\partial Q_1}{\partial \beta_k} = \sum_{i=1}^{n} \tau_{ik} \left(-\frac{\alpha_k}{\beta_k} + \frac{y_{i3}}{\beta_k^2} \right) = 0, \tag{A.6}$$

from which the following derivation is obtained:

$$\begin{aligned} \frac{\alpha_k \sum_{i=1}^{n} \tau_{ik}}{\beta_k} &= \frac{\sum_{i=1}^{n} \tau_{ik} y_{i3}}{\beta_k^2} \\ \implies \beta_k \alpha_k \sum_{i=1}^{n} \tau_{ik} &= \sum_{i=1}^{n} \tau_{ik} y_{i3}. \end{aligned}$$

The closed form solution of β_k at the (s+1)th iteration of the M-step is given as

$$\beta_k^{(s+1)} = \frac{\sum_{i=1}^{n} \tau_{ik}^{(s+1)} y_{i3}}{\alpha_k^{(s)} \sum_{i=1}^{n} \tau_{ik}^{(s+1)}}.$$

$$\frac{\partial Q_1}{\partial \alpha_k} = \sum_{i=1}^{n} \tau_{ik}(-\log(\beta_k) - \frac{\Gamma'(\alpha_k)}{\Gamma(\alpha_k)} + \log(y_{i3})). \tag{A.7}$$

Let $\psi(\alpha) = \frac{\Gamma'(\alpha)}{\Gamma(\alpha)}$. Then

$$\frac{\partial Q_1}{\partial \alpha_k} = \sum_{i=1}^{n} \tau_{ik}(-\log(\beta_k) - \psi(\alpha_k) + \log(y_{i3})) = 0$$
$$\implies \psi(\alpha_k) = \frac{\sum_{i=1}^{n} \tau_{ik}(-\log(\beta_k) + \log(y_{i3}))}{\sum_{i=1}^{n} \tau_{ik}};$$

Using the solution for $\beta_k^{(s+1)}$, Equation A.7 can be rewritten as follows:

$$\frac{\partial Q_1}{\partial \alpha_k} = \sum_{i=1}^{n} \tau_{ik}(-\log(\frac{\sum_{i=1}^{n} \tau_{ik} y_{i3}}{\alpha_k \sum_{i=1}^{n} \tau_{ik}}) - \psi(\alpha_k) + \log(y_{i3})) = 0. \tag{A.8}$$

Since Equation A.8 cannot be solved in closed form, the maximum likelihood of $\alpha_k^{(s+1)}$ cannot be solved in close form, so Nelder-Mead method was used to obtain estimate for $\alpha_k^{(s+1)}$.

For π_k, it is known that $\sum_{k=1}^{K} \pi_k = 1$. Applying Lagrange multiplier gives

$$Q_2^* = \sum_{i=1}^{N} \tau_{ik}(\log \pi_k) + \lambda \left(\sum_{k=1}^{K} \pi_k - 1 \right)$$

The estimate of π_k at the (s+1)th iteration of the M-step is obtained through partial derivative of Q_2^* with respect to π_k.

$$\frac{\partial Q_2^*}{\partial \pi_k} = \frac{\sum_{i=1}^{N} \tau_{ik}}{\pi_k} - \lambda = 0,$$

from which we obtain

$$\pi_k = \frac{\sum_{i=1}^{N} \tau_{ik}}{\lambda}.$$

This can be rearranged to

$$\sum_{k=1}^{K} \pi_k = \sum_{k=1}^{K} \frac{\sum_{i=1}^{N} \tau_{ik}}{\lambda}$$
$$\implies 1 = \frac{\sum_{k=1}^{K} \sum_{i=1}^{N} \tau_{ik}}{\lambda}.$$

Since $\frac{\sum_{k=1}^{K}\sum_{i=1}^{N}\tau_{ik}}{\lambda} = 1$, $\lambda = \sum_{k=1}^{K}\sum_{i=1}^{N}\tau_{ik} = N$. Therefore, the estimate becomes

$$\pi_k^{(s+1)} = \frac{\sum_{i=1}^{N}\tau_{ik}^{(s+1)}}{N}.$$

A.2 INDEPENDENCE ASSUMPTION

Independence among variables in an observation was assumed to calculate the log-likelihood $\log L(\boldsymbol{\psi}|\boldsymbol{y}_i)$ as mentioned in Section 2.[19] We performed a simple test of correlation. We calculated pairwise Pearson correlation among variables and performed significance test. The sample Pearson correlation coefficient of $\boldsymbol{y}_1$ and $\boldsymbol{y}_2$ follows the following formula:

$$r_{\boldsymbol{y}_1\boldsymbol{y}_2} = \frac{\sum_{i=1}^{N}(y_{i1}-\bar{y_1})(y_{i2}-\bar{y_2})}{\sqrt{\sum_{i=1}^{N}(y_{i1}-\bar{y_1})^2}\sqrt{\sum_{i=1}^{N}(y_{i2}-\bar{y_2})^2}}, \tag{A.9}$$

where $\bar{y_1}$ and $\bar{y_2}$ represent sample means of $\boldsymbol{y}_1$ and $\boldsymbol{y}_2$, respectively.[15] We performed paired Student's t-test, to determine if correlation coefficients were significantly different from zero, which would imply linear relationship among variables. T-statistic takes the following form:

$$t = r\sqrt{\frac{N-2}{1-r^2}}, \tag{A.10}$$

where r represents the correlation coefficient calculated using Equation A.9, N represents sample size, and $N-2$ corresponds to the degrees of freedom. However, with large sample sizes, Student's t-distribution is functionally equivalent to normal distribution.[9]

A.3 INDEPENDENCE ASSUMPTION BY CLASS

In Section 3, student's t-test was performed over the final data set with $4,841$ observations. In this subsection, we present output of the tests performed over observations in each of the three classes to examine if independence assumption can be met if the test is performed separately by class membership.

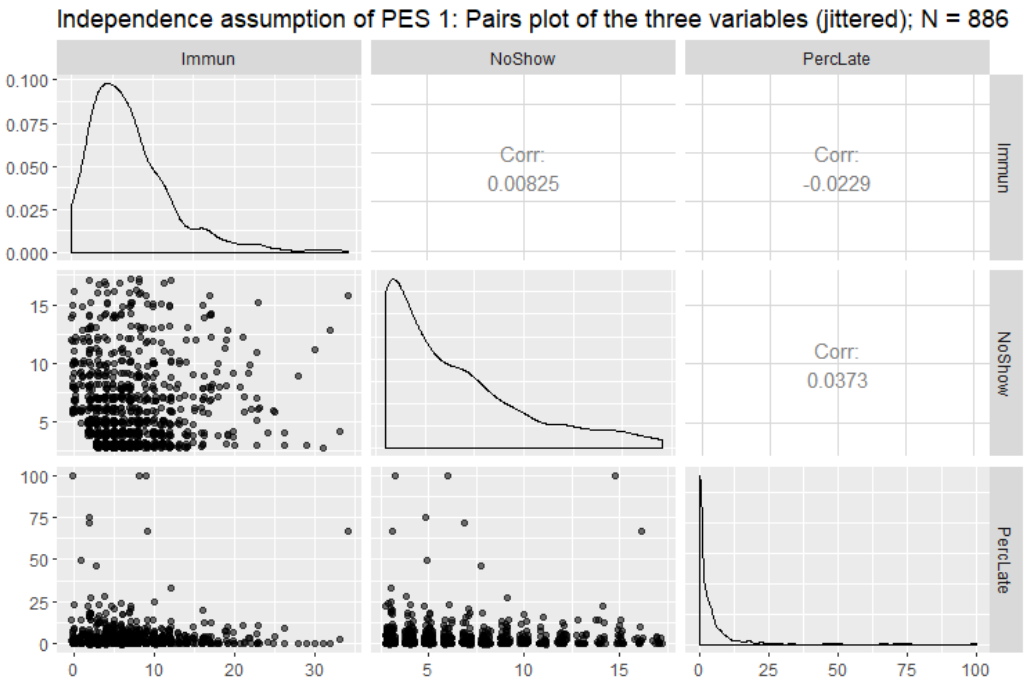

Figure 13: Pairs plots of the three measurements. $N = 4,841$

Table 11: Linear correlations and p-value estimates of observations belonging to PES 1

Comparisons	Correlation	t-statitic	P-value
Immun vs. NoShow	0.0082	0.2452	0.8063
Immun vs. PercLate	−0.0229	−0.6801	0.4966
NoShow vs. PercLate	0.0373	1.1112	0.2668

In contrast to the results we obtained without separating by classes, all pairwise correlations were found to be not significantly different from zero when only PES 1 was selected. However, a quick glance over the Figure 13 indicates that non-linear relationship may exist among the three measurements.

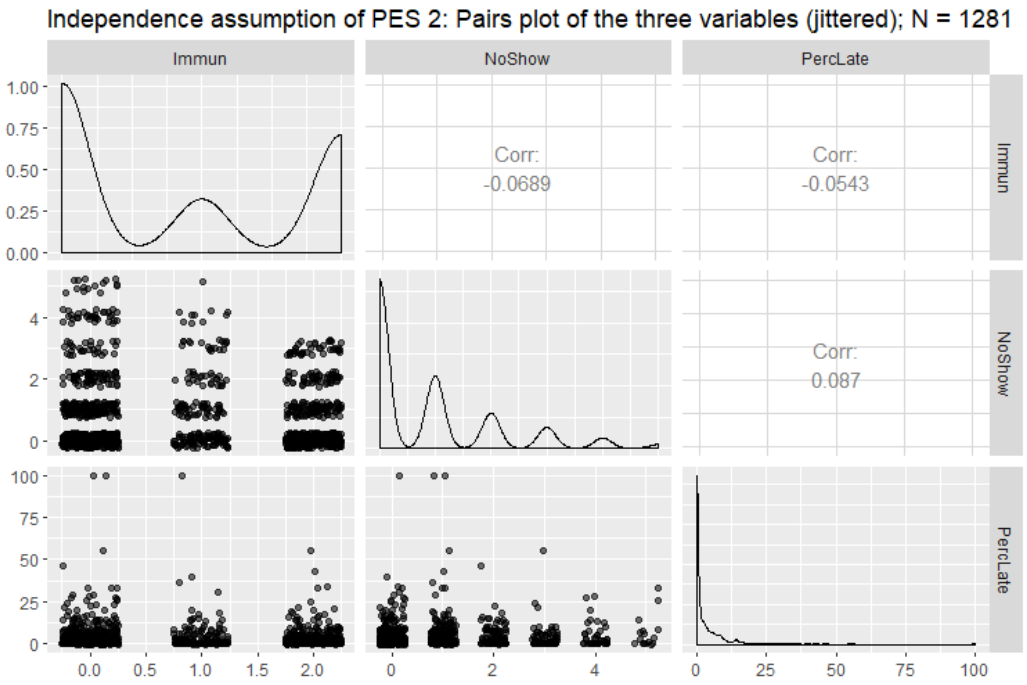

Figure 14: Pairs plots of the three measurements. $N = 4,841$

Table 12: Linear correlations and p-value estimates of observations belonging to PES 2

Comparisons	Correlation	t-statitic	P-value
Immun vs. NoShow	−0.0689	−2.4693	**0.0137**
Immun vs. PercLate	−0.0543	−1.9461	0.0519
NoShow vs. PercLate	0.0870	3.1239	**0.0018**

The result of the test for PES 2 was more similar to that obtained from Section 3 than to that from PES 1; the correlations were small but found to be significantly different from zero for two of the pairs - *Immun & NoShow* and *NoShow & PercLate*.

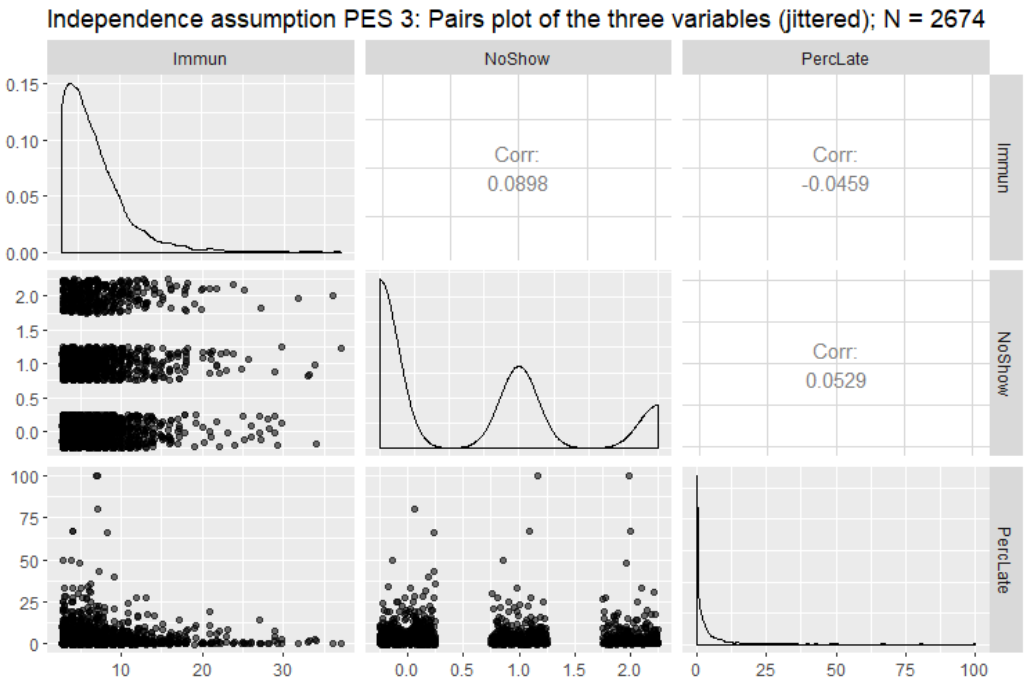

Figure 15: Pairs plots of the three measurements. $N = 4,841$

Table 13: Linear correlations and p-value estimates of observations belonging to PES 3

Comparisons	Correlation	t-statitic	P-value
Immun vs. NoShow	0.0898	4.6630	$<$ **0.0001**
Immun vs. PercLate	-0.0459	-2.3750	**0.0176**
NoShow vs. PercLate	0.0529	2.7373	**0.0062**

When only observations in PES 3 were selected, all three comparisons returned with correlations that were significantly different from zero. Overall assumption of independence does not appear to have met after separating observations by class membership. However, other multivariate independence tests may need to be considered to find a conclusive result. The same caveat from Section 3 holds: depending on the perspective of the readers, despite low p-values, assumptions of independence may be reasonable due to the estimated correlations being relatively low.

A.4 ESTIMATION OF PARAMETERS

In Section 2, we discussed obtaining estimates for $\boldsymbol{\theta}^{(s+1)}$ using EM Algorithm. The final calculated estimates of the $\boldsymbol{\theta}$ of the three different models are listed in Table 14, and Table 15. These values were used to generate log-likelihoods and histograms of expected count of the three measurements, as seen in Figures 5 and 8.

Table 14: Estimated values of unknown parameters of model $GGGK5$. Class membership before reordering

	$K=1$	$K=2$	$K=3$	$K=4$	$K=5$
$\widehat{\pi}_k$	0.1429	0.1489	0.1993	0.1277	0.3811
N_k	692	721	965	618	1845
$\widehat{\alpha}_{k,Immun}$	0.0599	5.9724	7.1893	0.1366	4.7561
$\widehat{\alpha}_{k,NoShow}$	0.0865	0.0755	0.0687	0.1050	0.1017
$\widehat{\alpha}_{k,PercLate}$	0.0681	2.8204	0.0562	1.9494	0.0619
$\widehat{\beta}_{k,Immun}$	0.1382	0.8844	2.5877	0.0514	0.5293
$\widehat{\beta}_{k,NoShow}$	0.0515	0.0866	0.1239	0.0463	0.0412
$\widehat{\beta}_{k,PercLate}$	0.0832	0.6391	0.2142	0.1382	0.1082

Table 15: Estimated values of unknown parameters of model $GGGK3$. Class membership AFTER reordering

	$K=1$	$K=2$	$K=3$
$\widehat{\pi}_k$	0.6602	0.2913	0.0485
N_k	3196	1410	235
$\widehat{\alpha}_{k,Immun}$	0.1068	0.0930	0.0681
$\widehat{\alpha}_{k,NoShow}$	0.0835	16.8643	3.5288
$\widehat{\alpha}_{k,PercLate}$	0.0995	0.1334	0.0796
$\widehat{\beta}_{k,Immun}$	0.0161	0.0333	0.0583
$\widehat{\beta}_{k,NoShow}$	0.0497	0.8729	0.5886
$\widehat{\beta}_{k,PercLate}$	0.0388	0.0418	0.064

A.5 MODEL SELECTION

Figures 16 and 17 are residual plots of the two discrete measurements, *Immun* and *NoShow*, of the three models. They correspond to the alternative way of visualizing Figures 5 and 8. For each model, there are two sets of residual plots -

absolute and weighted. Absolute residuals are calculated by a simple difference, $AbsResidual_v = Observed_v - Expected_v$, were $v = 0, 1, ...$ Relative residuals were calculated by $RelResidual_v = \frac{Observed_v - Expected_v}{Observed_v + Expected_v}$ in order to supplement the shortcomings of the absolute residuals by exploring how the observed and expected counts differed as a ratio. Therefore, in relative residuals, for a certain value, if the number of observed was 15 and expected was 5, the weighted residuals would be $\frac{15-5}{15+5} = \frac{10}{20} = 0.5$. For larger values, where either observed or expected counts were scarce or nonexistent, the relative residuals took the values close or equal to 1 or -1.

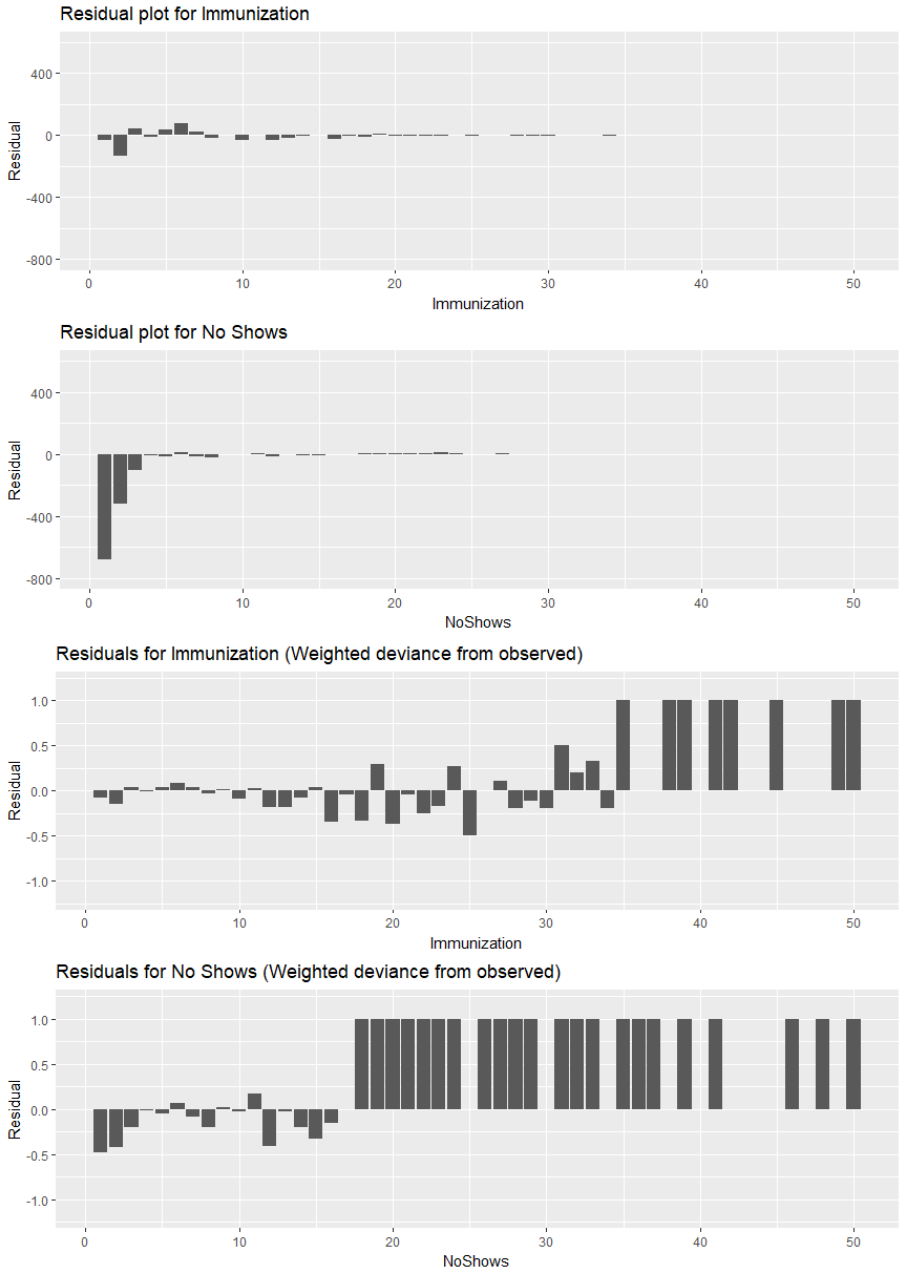

Figure 16: Residual plots of *Immun* and *NoShow* of the model $GGGK5$

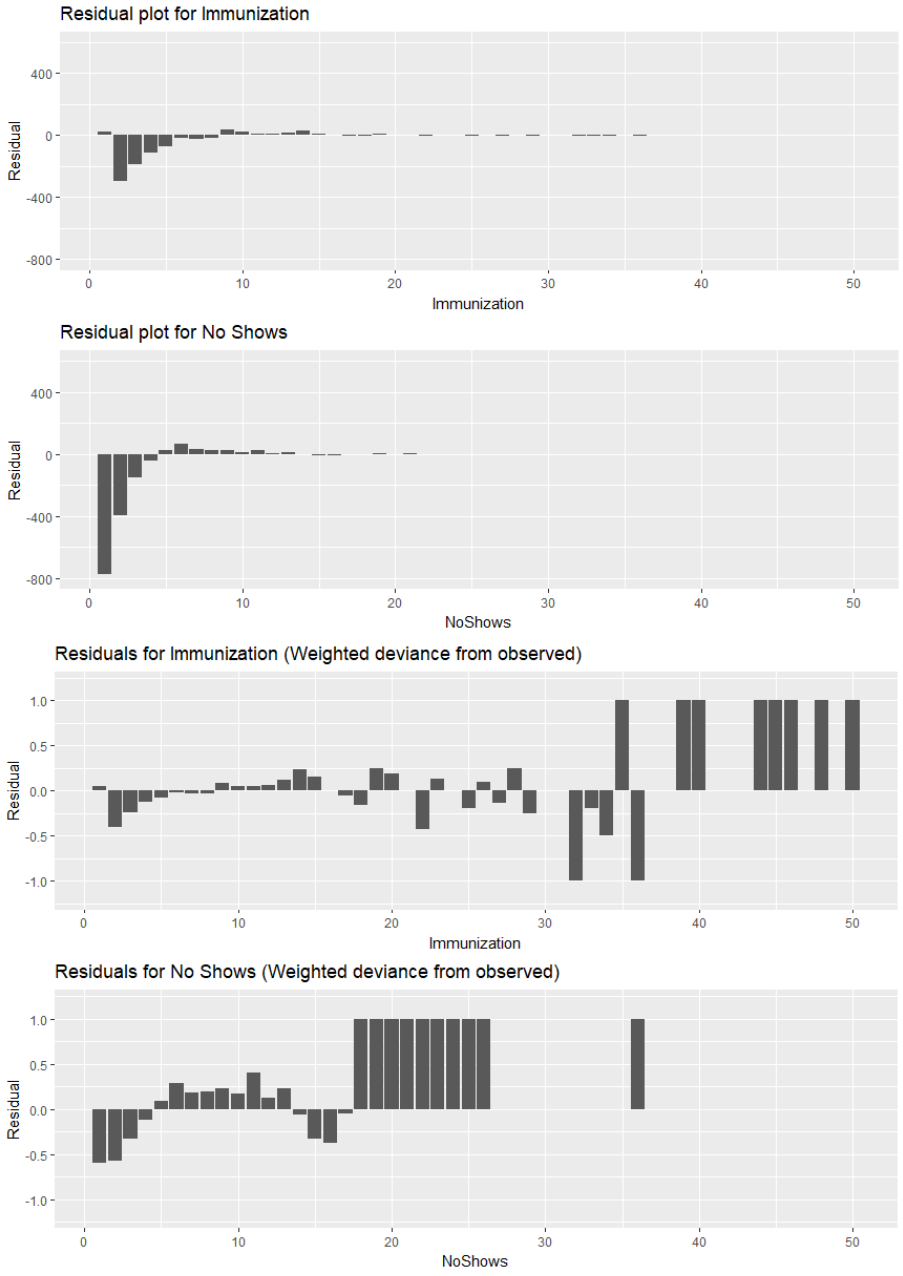

Figure 17: Residual plots of *Immun* and *NoShow* of the model $GGGK3$

A.6 MULTIPLE LOGISTIC REGRESSION

The following four tables (Table 16-19) are the summary outputs from multiple logistic regression for the association between PES and the eight health-related outcomes controlled by patient characteristics and health behaviors.

Table 16: Multiple logistic regression for the association between Patient Engagement Score (PES) and health-related outcomes adjusted by patient characteristics and health behavior. Model used: $GGGK3$

	Systolic BP*			**Diastolic BP***		
Variable	Coefficient (SE)	p-value	Odds Ratio	Coefficient (SE)	p-value	Odds Ratio
(Intercept)	-3.511 (0.485)	**< 0.001**	0.030	-1.837 (0.653)	**0.005**	0.159
PES						
PES1	0.027 (0.139)	0.846	1.027	0.134 (0.204)	0.510	1.144
PES2**						
PES3	-0.177 (0.112)	0.114	0.838	0.020 (0.166)	0.904	1.020
Age	0.020 (0.006)	**< 0.001**	1.021	-0.032 (0.007)	**< 0.001**	0.969
Gender						
Female	-0.153 (0.098)	0.117	0.858	-0.631 (0.151)	**< 0.001**	0.532
Male**						
Marital Status						
Married	-0.163 (0.105)	0.121	0.850	-0.189 (0.155)	0.224	0.828
Not Married**						
Race						
Non-whites	0.217 (0.204)	0.288	1.242	0.473 (0.263)	0.072	1.604
Whites**						
Primary Insurance						
Govt	0.080 (0.130)	0.536	1.084	0.202 (0.210)	0.335	1.224
Non-Govt**						
BMI	0.026 (0.007)	**< 0.001**	1.026	0.037 (0.011)	**< 0.001**	1.038
Tobacco						
Non-smoker	-0.515 (0.126)	**< 0.001**	0.597	-0.150 (0.193)	0.435	0.861
Smoker**						
Alcohol						
No	0.058 (0.096)	0.543	1.060	-0.120 (0.145)	0.405	0.887
Yes**						
Chronic Condition	0.025 (0.027)	0.349	1.026	-0.039 (0.046)	0.403	0.962

N = 4,841

Table 17: Multiple logistic regression for the association between Patient Engagement Score (PES) and health-related outcomes adjusted by patient characteristics and health behavior. Model used: $GGGK3$.

		HDL*			**LDL***	
Variable	Coefficient (SE)	p-value	Odds Ratio	Coefficient (SE)	p-value	Odds Ratio
(Intercept)	-0.617 (0.320)	0.054	0.540	1.339 (0.303)	**< 0.001**	3.815
PES						
PES1	-0.211 (0.099)	0.033	0.810	-0.151 (0.093)	0.106	0.860
PES2**						
PES3	-0.206 (0.076)	**0.007**	0.814	-0.280 (0.072)	**< 0.001**	0.756
Age	-0.016 (0.004)	**< 0.001**	0.985	-0.017 (0.003)	**< 0.001**	0.983
Gender						
Female	-1.236 (0.070)	**< 0.001**	0.291	0.487 (0.062)	**< 0.001**	1.627
Male**						
Marital Status						
Married	-0.044 (0.075)	0.561	0.957	0.043 (0.069)	0.529	1.044
Not Married**						
Race						
Non-whites	-0.225 (0.153)	0.141	0.799	-0.095 (0.140)	0.499	0.910
Whites**						
Primary Insurance						
Govt	0.148 (0.095)	0.120	1.159	0.019 (0.090)	0.831	1.019
Non-Govt**						
BMI	0.040 (0.005)	**< 0.001**	1.040	-0.008 (0.005)	0.107	0.992
Tobacco						
Non-smoker	-0.539 (0.095)	**< 0.001**	0.583	-0.130 (0.091)	0.151	0.878
Smoker**						
Alcohol						
No	0.367 (0.066)	**< 0.001**	1.444	-0.090 (0.062)	0.148	0.914
Yes**						
Chronic Condition	0.117 (0.020)	**< 0.001**	1.125	-0.151 (0.021)	**< 0.001**	0.860

N = 4,841

Table 18: Multiple logistic regression for the association between Patient Engagement Score (PES) and health-related outcomes adjusted by patient characteristics and health behavior. Model used: $GGGK3$.

	ED Visit*			**A1C Level***		
Variable	Coefficient (SE)	p-value	Odds Ratio	Coefficient (SE)	p-value	Odds Ratio
(Intercept)	-1.051 (0.313)	**0.001**	0.350	0.261 (0.299)	0.382	1.298
PES						
PES1	1.122 (0.095)	**< 0.001**	3.071	-0.138 (0.092)	0.137	0.871
PES2**						
PES3	0.086 (0.077)	0.266	1.090	-0.153 (0.071)	0.032	0.859
Age	-0.010 (0.003)	**0.003**	0.990	-0.023 (0.003)	**< 0.001**	0.978
Gender						
Female	0.012 (0.065)	0.850	1.012	0.009 (0.062)	0.879	1.009
Male**						
Marital Status						
Married	-0.023 (0.072)	0.754	0.978	-0.059 (0.068)	0.389	0.943
Not Married**						
Race						
Non-whites	0.227 (0.142)	0.109	1.255	0.341 (0.140)	0.015	1.406
Whites**						
Primary Insurance						
Govt	0.152 (0.092)	0.098	1.164	-0.140 (0.089)	0.113	0.869
Non-Govt**						
BMI	-0.009 (0.005)	0.070	0.991	0.012 (0.005)	**0.009**	1.012
Tobacco						
Non-smoker	-0.134 (0.094)	0.153	0.874	-0.153 (0.090)	0.091	0.859
Smoker**						
Alcohol						
No	0.228 (0.064)	**< 0.001**	1.256	0.308 (0.061)	**< 0.001**	1.360
Yes**						
Chronic Condition	0.154 (0.020)	**< 0.001**	1.166	0.222 (0.020)	**< 0.001**	1.249

N = 4,841

Table 19: Multiple logistic regression for the association between Patient Engagement Score (PES) and health-related outcomes adjusted by patient characteristics and health behavior. Model used: $GGGK3$.

		eGFR*			**Hospitalization***	
Variable	Coefficient (SE)	p-value	Odds Ratio	Coefficient (SE)	p-value	Odds Ratio
(Intercept)	-2.838 (0.336)	**< 0.001**	0.059	-2.041 (0.351)	**< 0.001**	0.130
PES						
PES1	-0.059 (0.103)	0.567	0.943	1.255 (0.103)	**< 0.001**	3.508
PES2**						
PES3	0.024 (0.081)	0.765	1.024	0.188 (0.089)	0.035	1.206
Age	0.064 (0.004)	**< 0.001**	1.066	0.007 (0.004)	0.076	1.007
Gender						
Female	0.582 (0.072)	**< 0.001**	1.790	-0.188 (0.072)	**0.009**	0.829
Male**						
Marital Status						
Married	0.007 (0.079)	0.934	1.007	-0.033 (0.079)	0.677	0.967
Not Married**						
Race						
Non-whites	-0.238 (0.149)	0.110	0.788	-0.186 (0.159)	0.242	0.830
Whites**						
Primary Insurance						
Govt	-0.027 (0.110)	0.803	0.973	0.295 (0.096)	**0.002**	1.344
Non-Govt**						
BMI	-0.014 (0.005)	**0.008**	0.986	-0.011 (0.006)	0.054	0.989
Tobacco						
Non-smoker	0.459 (0.096)	**< 0.001**	1.582	0.000 (0.104)	0.998	1.000
Smoker**						
Alcohol						
No	0.037 (0.070)	0.599	1.038	0.228 (0.071)	**0.001**	1.256
Yes**						
Chronic Condition	0.047 (0.023)	0.039	1.048	0.183 (0.020)	**< 0.001**	1.201

N = 4,841

REFERENCES

[1] Marjory Charlot, Michael R Winter, Howard Cabral, Michael S Wolf, Laura M Curtis, Amresh Hanchate, and Michael Paasche-Orlow. Patient activation mediates health literacy associated with hospital utilization among whites. *Health Literacy Research and Practice*, 1(3):128–135, 2017.

[2] Ronald Dendere, Christine Slade, Andrew Burton-Jones, Clair Sullivan, Andrew Staib, and Monika Janda. Patient portals facilitating engagement with inpatient electronic medical records: a systematic review. *Journal of Medical Internet Research*, 21(3):1, 2019.

[3] Ülkü Erişoğlu, Murat Erişoğlu, and Hamza Erol. A mixture model of two different distributions approach to the analysis of heterogeneous survival data. *International Journal of Computational and Mathematical Sciences*, 5(2), 2011.

[4] Andrew Gelman and Jennifer Hill. *Data analysis using regression and multilevel/hierarchical models*, page 46. Cambridge University Press, 1st edition, 2006.

[5] Jacques A Hagenaars and Allan L McCutcheon. *Applied latent class analysis*. Cambridge University Press, 2002.

[6] Insignia Health. Patient activation measure (pam), 2018.

[7] Judith H Hibbard, Eldon R Mahoney, Jean Stockard, and Martin Tusler. Development and testing of a short form of the patient activation measure. *Health Services Research*, 40(6 Pt 1):1918–1930, 2005.

[8] Judith H Hibbard, Jean Stockard, Eldon R Mahoney, and Martin Tusler. Development of the patient activation measure (pam): Conceptualizing

and measuring activation in patients and consumers. *Health Services Research*, 39(4 Pt 1):10051026, 2004.

[9] Maurice G. Kendall and Alan Stuart. *The Advanced Theory of Statistics, Vol. 2: Inference and Relationship*, page Section 31.19. Griffin, 3rd edition, 1973.

[10] Ariel Linden. Estimating measurement error of the patient activation measure for respondents with partially missing data. *BioMed Research International*, 2015(270168):1–7, 2015.

[11] Jillian A Macklin, Natalie Djihanian, Tieghan Killackey, and Jane MacIver. Engaging patients in care (epic): a framework for heart function and heart transplant - specific patient engagement. *CJC Open*, 1:43–46, 2019.

[12] Ranjan Maitra. Initializing partition-optimization algorithms. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 6:144–157, 2009.

[13] Semhar Michael and Volodymyr Melnykov. An effective strategy for initializing the em algorithm in finite mixture models. *Adv Data Anal Classif*, 10:563–583, 2016.

[14] Surachat Ngorsuraches, Patricia Da Rosa, Xijin Ge, Gemechis Djira, Semhar Michael, and Haward Wey. Php6 - patient engagement as a predictor for health outcomes and costs in multiple chronic conditions. *Value in Health*, 21(1):S88–S89, 2018.

[15] Karl Pearson. Notes on regression and inheritance in the case of two parents. *Proceedings of the Royal Society of London*, 58:240–242, 1895.

[16] Frank Picard. An introduction to mixture models, 2007.

[17] Jany Rademakers, Helle Terkildsen Maindal, Aslak Steinsbekk, Jochen Gensichen, Katja Brenk-Franz, and Michelle Hendriks. Patient activation in europe: an international comparison of psychometric properties and patients scores on the short form patient activation measure (pam-13). *BMC health services research*, 16(1):570, 2016.

[18] Samira Abbasgholizadeh Rahimi, Herv Tchala Vignon Zomahoun, and France Lgar. Patient engagement and its evaluation tools - current challenges and future directions. *International Journal of Health Policy and Management*, 8(6):378–380, 2019.

[19] Randy Read. Likelihood: theory and application to structure refinement, 2001.

[20] Gideon Schwarz. Estimating the dimension of a model. *The Annals of Statistics*, 6(2):461–464, 1978.