CROSS-VALIDATION IN GROUP-BASED LATENT TRAJECTORY MODELING WHEN ASSUMING A CENSORED NORMAL MODEL

by

Megan M. Marron

B.S., Rochester Institute of Technology, 2011

Submitted to the Graduate Faculty of

the Graduate School of Public Health in partial fulfillment

of the requirements for the degree of

Master of Science

University of Pittsburgh

2014

UNIVERSITY OF PITTSBURGH

GRADUATE SCHOOL OF PUBLIC HEALTH

This thesis was presented

by

Megan M. Marron

It was defended on

December 5, 2014

and approved by

Francis E. Lotrich, PhD, M.D., Associate Professor, Department of Psychiatry, Western Psychiatry Institute and Clinic, University of Pittsburgh School of Medicine

Ada O. Youk, PhD, Assistant Professor, Department of Biostatistics, Graduate School of Public Health, University of Pittsburgh

Thesis Director: Stewart J. Anderson, PhD, Professor, Department of Biostatistics, Graduate School of Public Health, University of Pittsburgh

Stewart J. Anderson, PhD

CROSS-VALIDATION IN GROUP-BASED LATENT TRAJECTORY MODELING WHEN ASSUMING A CENSORED NORMAL MODEL

Megan M. Marron, M.S.

University of Pittsburgh, 2014

ABSTRACT

Group-based latent trajectory modeling (GBLTM) is a relatively recent addition to methodology for analyzing repeated measures of a variable over time. It is implemented using SAS procedure TRAJ for a zero-inflated Poisson (ZIP) model, censored normal (CNORM) model, and logistic model. Cross-validation (CV) in GBLTM is used as an alternative tool to the Bayesian Information Criterion (BIC) for determining the optimal number of distinct latent subgroups in a sample. CV in GBLTM when assuming a ZIP model is implemented using the crimCV package in R. In this thesis, the use of CV in GBLTM is furthered by applying it when assuming a CNORM model and examining the consistency of results when considering multiple types of CV. This method is applied to a Hepatitis C (HCV) study to determine whether patterns of depressive symptoms in HCV patients treated with interferon-based therapy form clinically meaningful distinct subgroups. When applied to the HCV study, CV was a conservative approach to model selection compared with BIC; CV suggested a two-group model, whereas BIC suggested a five-group model. However, when visually examining the data, a three-group model appeared to capture the heterogeneity in the HCV sample best. Therefore, BIC and CV should not be used alone to determine the optimal number of distinct latent subgroups in a sample, but rather used to make an educated judgment on the number of subgroups that describes the heterogeneity best. Whether or not CV is truly a conservative approach to model selection compared with BIC is still unknown, CV and BIC should be further explored using other datasets and simulations. The public health significance of this thesis is exploring statistical tools used for determining the optimal number of distinct latent subgroups in GBLTM, where knowledge of the factors that predispose individuals to less favorable trajectory groups, can lead to targeted preventions.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

PREFACE

I would like to express my sincere gratitude to all my committee members: Dr. Francis E. Lotrich for introducing me to such an important public health topic, for allowing me to use his study to produce this methodological thesis, and ultimately for giving me the opportunity to take my academic career to the next level; Dr. Ada O. Youk for her practical guidance and support; and especially Dr. Stewart J. Anderson for his expert methodological knowledge, teaching, and mentorship throughout my degree.

I would also like to thank Dr. Charles F. Reynolds III and his Advanced Center for Intervention and Services Research for Late-Life Mood Disorders, especially Amy Begley for her essential role in my growth as a biostatistician, as well as her irreplaceable friendship during my time as a Graduate Student Researcher. I would like to thank the friends and classmates that I have met at the University of Pittsburgh, especially Stephen F. Smagula for providing helpful discussion and realistic insight into my statistical work.

Lastly, I would like to thank my mom and dad, my brother John, and my sisters Courtney and Chelsea, for their love and support throughout all of my endeavors. Thank you for helping me become the person I am today.

1.0 INTRODUCTION

1.1 OVERVIEW

Group-based latent trajectory modeling (GBLTM) is a relatively new addition to the methodology of analyzing longitudinal data. GBLTM is useful when one is interested in determining whether there exist distinct subgroups in a sample, where each subgroup consists of individuals who share similar patterns of repeated measures over time. The Bayesian Information Criterion (BIC) is used for model selection in GBLTM to determine the optimal number of distinct subgroups existing in a sample. However, BIC sometimes suggests an unreasonable amount of subgroups [1-6].

Cross-validation (CV) is a statistical method of evaluating and comparing different models by dividing available data into two sets used for constructing a model and to validate the model that was constructed [7]. CV can be used to estimate the generalizability of a model's performance and is thus an alternative to BIC for model selection. CV is implemented in GBLTM for determining the optimal number of latent subgroups in a zero-inflated Poisson (ZIP) model using an R package, crimCV [8].

In this thesis, the use of CV in GBLTM was furthered by applying it when assuming a censored normal (CNORM) model. Different types of CV were examined to determine if each type suggests a consistent number of latent subgroups and how they compare with BIC. We used

CV in GBLTM to evaluate longitudinal data from a hepatitis C (HCV) study, with the aim of determining the number of clinically meaningful distinct subgroups of patients who share similar patterns of change in depressive symptoms over the course of treatment.

1.2 GROUP-BASED LATENT TRAJECTORY MODELING (GBLTM)

There are three main branches of group-based methodology used to analyze repeated measures of a variable over time, one of them being group-based latent trajectory modeling (GBLTM) [9]. The others are hierarchical modeling [10-12] and growth curve analysis [13-16]. All three methodologies strive to model individual-level heterogeneity in longitudinal data. Hierarchical modeling uses the strategy of a random coefficient model to capture individual variation in repeated measures over time, whereas growth curve analysis uses the method of a covariance structure. Both of these methods assume the parameters of a trajectory follow the same distribution, meaning the population distribution of trajectories varies across individuals continuously and can be modeled in most cases using a multivariate normal distribution [9]. GBLTM takes a different approach by using a semiparametric multinomial modeling technique, making the assumption that there are distinct groups of trajectories that exist due to underlying latent traits. A latent trait is an unobservable variable that is thought to cause a pattern in observed variables. GBLTM was created to compliment hierarchical modeling and growth curve analysis [17]. Note that, while we focus specifically on the strategy implemented using SAS TRAJ procedure (Proc TRAJ), GBLTM can also be employed in a Structural Equation Modeling (SEM) framework.

GBLTM partitions a cohort into subgroups based on repeated measures of a variable over time and baseline risk factors [17-19]. This enables one to classify individuals based on latent traits that can change over time. Nagin and Land first introduced the concept in their 1993 manuscript entitled *“Age, Criminal Careers, and Population Heterogeneity: Specification and Estimation of a Nonparametric, Mixed Poisson Model”*, where they used a semiparametric estimation procedure in order to identify distinct groups of criminals who shared similar patterns of offending over time [20]. In 1996 Land, McCall, and Nagin furthered the mixture model approach [21] and in 1999 Roeder, Lynch, and Nagin were able to allow for the uncertainty of latent class membership, as well as incorporate time stable covariates [19]. The statistical theory continued to develop [6, 22] and in 2001, Jones, Nagin, and Roeder created Proc TRAJ to implement GBLTM for three different models: ZIP model, CNORM model, and logistic (LOGIT) model [17].

GBLTM is a data-driven procedure that describes the heterogeneity existing in populations that may be difficult to identify a priori. An advantage of this technique is that rather than examining the population-average change, distinct subgroups are identified empirically, where each subgroup consists of subjects who share similar patterns of repeated measures over time. In addition, risk factors can then be examined to explore whether demographic and clinical variables of interest measured at baseline are associated with specific subgroups.

A model selection process using BIC is already well established in GBLTM [19, 23], where the first step is determining the optimal number of distinct subgroups using BIC. BIC [24] is a statistical tool used to evaluate model fit, which is based on the log-likelihood evaluated at the maximum likelihood estimate (MLE) with an additional penalty term for the number of parameters added. The model with the number of latent subgroups that maximizes the BIC should be chosen, as long as the difference in BIC between the next best model is meaningful. The log Bayes factor approximation is a useful tool for determining a meaningful difference in BIC. It is classified into four groups, as the amount of evidence against a simpler model: “Not worth mentioning”, “Positive”, “Strong”, and “Very Strong”. Unfortunately, programs such as Proc TRAJ and Mplus sometimes compute a BIC that suggests an unreasonable amount of groups [8], where the model will contain a redundant group and/or a group consisting of less than 5% of the sample [1-6].

1.3 CROSS-VALIDATION (CV)

Cross-validation (CV) is a statistical method that can be used in model selection. CV evaluates and compares different models by dividing available data into two sets, which are used for constructing a model (training set) and validating the model (validation set) [7]. Statistical analyses often aim to learn a model from available data, but this may be difficult because the model may demonstrate adequate prediction capability on the available data but fail to predict future unseen data. To address this issue, CV is a method which estimates the generalizability of a model's performance. The idea of CV originated in 1931 [25], and the idea of CV that is similar to today's k-fold CV was introduced in 1968 [26].

In 2011, Nielsen et al. proposed CV as an alternative to BIC for determining the number of latent subgroups in GBLTM when assuming a zero-inflated Poisson (ZIP) model [8]. They implement leave-one-out CV (LOOCV) in R package crimCV for determining the optimal number of latent subgroups in a ZIP model [8]. LOOCV is also known as n-fold CV and jackknife resampling. Other types of CV include re-substitution validation, hold-out validation, k-fold CV, and repeated k-fold CV. This thesis focuses on five-fold CV, five by two-fold, and LOOCV.

1.4 CURRENT PROJECT

As discussed above, latent traits that are thought to cause patterns in repeated measures over time may be difficult to identify a priori. When using population-average techniques in trajectory modeling (such as hierarchal models and growth curve analysis), it is often unclear whether an association found represents an overall trend or if it is driven by a specific subgroup of the sample. GBLTM makes it possible to explore whether such associations are in fact driven by a specific subgroup. Knowledge of factors that predispose individuals to less favorable trajectories, can lead to targeted prevention. For this thesis, multiple types of CV are applied to GBLTM when assuming a censored normal model. A SAS macro is developed to apply CV in GBLTM while using Proc TRAJ. Therefore the novel aspect presented here are: (1) use of CV in GBLTM for CNORM models as an alternative to BIC in model selection, (2) comparison of the models' consistency using multiple types of CV, and (3) interpreting the differences between model selection criteria.

This method is applied to a hepatitis C (HCV) study, to determine how many distinct subgroups of change in depressive symptoms exist during the course of interferon-based therapy. The optimal number of latent groups in the HCV study was first selected using BIC and again using multiple types of CV.

1.5 PUBLIC HEALTH SIGNIFICANCE

GBLTM is a statistical tool that enables one to classify the across-patient heterogeneity to get a better picture of the prognosis of different latent subgroups that exist within the population. Population average techniques, such as mixed models which pre-specify subgroups, are not always an effective tool to use when dealing with highly diverse patient populations because they assume that each fixed population follows the same distribution. GBLTM allows for each latent subgroup to follow a different distribution, making it possible to sort through longitudinal heterogeneity in a sample to allow for more person centered models. With a growing interest for personalized medicine, it is likely that GBLTM will continue to become increasingly popular.

There is a need for a more rigorous method for determining the true number of latent subgroups in a population. This thesis contributes to public health research by exploring different statistical tools used for determining the optimal number of latent subgroups in GBLTM. We test how CV may serve as an alternative tool to BIC for model selection in GBLTM. It is important to be able to find the best, most parsimonious model so that the number of latent subgroups fully describes the underlying heterogeneity in a population. Once the optimal number of groups is found, associations between baseline risk factors and group membership can be examined in order to discover why some subjects experience less favorable trajectories compared to others, or what causes subjects who are assigned to a group with a favorable trajectory to be so resilient throughout a study. It is important to be able to identify the number of latent subgroups based on an objective criterion, to support well informed targeted prevention techniques.

2.0 STATISTICAL METHODS

2.1 ESTIMATION OF GROUP-BASED TRAJECTORIES

A brief overview of the theory behind GBLTM is given, however much more detail can be found elsewhere [6, 19-22].

Let Y_{ij} be the longitudinal outcome of interest for the i^{th} subject at the j^{th} time point, where $i = 1, \dots, N$ subjects and $j = 1, \dots, J_i$ time points. Let $k = 2, \dots, K$ be the number of latent subgroups that exists in a population, where each subgroup consists of subjects who share similar measurements of the outcome over time. Let C_i be the unobservable latent variable of the i^{th} subject, where the latent variable is assumed to be the underlying reason to why a subject experienced a certain pattern in repeated measures of the outcome over time. Let z_i denote the i^{th} subject's measurements for time-stable covariates, also known as baseline risk factors. The model is then developed as a standard finite mixture model with k-components constructed using the following marginal density function:

$$f(y_{ij}) = \sum_{k=1}^{K} Pr(C_i = k | Z_i = z_i) Pr(Y_{ij} = y_{ij} | C_i = k),$$

where $Pr(C_i = k | Z_i = z_i)$ is the posterior probability of belonging to group k given measured baseline risk factors and $Pr(Y_{ij} = y_{ij} | C_i = k)$ is the probability of repeated measures of the outcome of interest over time given a subject belongs to the k^{th} latent subgroup of a population.

The posterior probability is modeled as a K-outcome logit function using multinomial logistic regression to illustrate the relationship between baseline risk factors and k latent subgroups. The posterior probability of belonging to the k^{th} group given certain baseline risk factors is calculated based off of Bayes' Theorem using the following equation:

$$\pi_i^k = Pr(C_i = k | Z_i = z_i) = \frac{exp\{\theta_k + \gamma_k' z_i\}}{\sum_{l=1}^{K} exp\{\theta_l + \gamma_l' z_i\}},$$

Where $\theta = (\theta_1, \dots, \theta_k)$ is the group effect and $\gamma = (\gamma_1, \dots, \gamma_k)$ is the additional group effect given baseline risk factors. The title posterior probability was given because these probabilities are computed after fitting a model, using the model's estimated coefficients [23].

The main focus of GBLTM is to learn about the relationship between baseline risk factors and latent subgroups, but to do this longitudinal data, Y_{ij}, has to be used in order to understand the unobservable latent trait involved in group membership [19]. There are three different models developed in Proc TRAJ to model the probability of observing a certain pattern in repeated measures of the outcome over time given the subject belongs to the k^{th} latent subgroup: ZIP model, CNORM model, and LOGIT model. This thesis will focus on the CNORM model.

The CNORM model is particularly useful for psychometric scale data, which addresses the problem of clustering at the minimum and maximum of the scale, also known as floor and ceiling effect. A ceiling effect is when a subject starts off with very high symptoms and cannot get a measurement much higher than that at their next follow-up visit because they were already near the maximum of the scale to start off with. Therefore, the rate of change of the individual’s trajectory may not be an accurate representation of their true change in symptoms. A floor effect is the same idea, except a subject is experiencing no symptoms, so their symptom count cannot decrease any more. A distribution that allows for censoring at the minimum and maximum of a scale addresses this issue of subjects clustering at low or high symptoms [17]. The likelihood of observing a specific pattern of repeated measures over time in subject i when they belong to latent subgroup k is:

$$Pr\left(Y_{ij} = y_{ij} | C_i = k\right)$$
$$= \prod_{y_{ij}=Min} \Phi\left[\frac{Min - \mu_{ijk}}{\sigma}\right] \prod_{Min<y_{ij}<Max} \phi\left[\frac{y_{ij} - \mu_{ijk}}{\sigma}\right] \prod_{y_{ij}=Max} 1 - \Phi\left[\frac{Max - \mu_{ijk}}{\sigma}\right],$$

where $\mu_{ijk} = \beta_{0k} + \beta_{1k} time_{ij} + \beta_{2k} time_{ij}^2 + \cdots + \beta_{pk} time_{ij}^p + \lambda w_{ij}$. The polynomial degree for each trajectory is denoted as p, where if μ consists of $p = 0,1,2$ then the trajectory is a second-degree polynomial, also known as a quadratic curve. Along with time-stable covariates, GBLTM can also incorporate time-varying covariates; the measurement of time-varying covariates for the i^{th} subject at the j^{th} time point is denoted as w_{ij} and λ is the effect of the time-varying covariate on trajectory mean. This thesis focuses only on time-stable covariates and we consider up to a second-degree polynomial, which reduces the expected trajectory for the i^{th} subject in the k^{th} latent subgroup to $\mu_{ijk} = \beta_{0k} + \beta_{1k} time_{ij} + \beta_{2k} time_{ij}^2$.

2.1.1 Computing the Bayesian Information Criterion (BIC)

BIC is recommended for model selection in many different methods of modeling. This includes, determining the number of groups underlying the method of GBLTM [23]. BIC is a tool similar to the likelihood ratio test; however BIC can be used with models that are not nested. BIC also has the advantage of a penalty term for extra parameters added. The main reason BIC is used in GBLTM instead of the likelihood ratio test is because the null hypothesis of choosing a smaller number of groups versus the alternative hypothesis of a larger number of groups is on the boundary of the parameter space. When this occurs, standard asymptotic results [27] no longer hold, and hence the change in BIC [28] is used instead. When this issue is ignored, the likelihood ratio test tends to favor the alternative hypothesis of a more complex model [29].

BIC is calculated for each model by summing two portions: one involving the likelihood and the other involving a penalty term for the number of parameters added to the model:

$$BIC = log\left[L\left(\hat{\theta}\right)\right] + 0.5qlog(n).$$

The log-likelihood evaluated at the MLE is denoted as $log\left[L\left(\hat{\theta}\right)\right]$. This portion of the BIC equation continues to improve as the number of latent subgroups increases. This term is similar to R^2 in linear regression because R^2 can only increase with the addition of another parameter [23]. However, the second half of the BIC equation counterbalances this effect, where q is the number of parameters added to the model and n is the sample size. The polynomial degree of each trajectory and the number of latent subgroups both contribute to the number of parameters in a model.

When comparing two models in GBLTM, the model with the largest BIC is chosen as long as the difference in BIC between the two models is meaningful. The Bayes factor is a useful statistic for determining whether the difference between BIC from two models is meaningful. It measures the posterior odds of the more complex model being the correct model for the available data compared to the simpler model [23]:

$$Bayes\ Factor = \frac{Pr(Complex\ Model\ is\ Correct)}{Pr(Simpler\ Model\ is\ Correct)}.$$

A Bayes factor of five is interpreted as the simpler model is five times more likely than the complex model, which is noted as moderate evidence for the simpler model [30]. However, calculating the Bayes factor is very difficult. In GBLTM, this is avoided by using the log Bayes factor approximation [9, 17]:

$$2\log_e(B_{10}) \approx 2(\Delta\text{BIC}).$$

The approximation is simply twice the change in BIC, where the change is the BIC from the more complex model minus the BIC from the simpler model:

$$\Delta\text{BIC} = \text{BIC}_{\text{complex}\,model} - BIC_{Simpler\ model}.$$

The log Bayes factor approximation is a way to explain the amount of evidence supporting a more complex model and is thus an easier way to assess whether the change in BIC between two models is meaningful. The amount of evidence against the null hypothesis of a simpler model can be categorized into four groups: “Not worth mentioning”, “Positive”, “Strong”, and “Very Strong” evidence against a simpler model (Table 1). This allows one to easily determine a meaningful change [17].

Table 1. Guideline of a meaningful change in BIC

Log Bayes factor approximation: $2(\Delta BIC)$	Evidence against the null hypothesis: H_0: simpler model (smaller number of groups)
<2	Not worth mentioning
2 to <6	Positive
6 to <10	Strong
≥10	Very strong

Note: adapted from [17]

2.1.2 Issues with BIC in GBLTM

The MLEs involved in GBLTM cannot be derived because their respective equations are not in closed-form. Therefore, Proc TRAJ uses EM algorithm to search for the parameter estimate that maximizes the likelihood [23]. Even when the EM algorithm successfully finds the correct MLEs, the BIC may still suggest an unreasonable amount of groups [1-6, 8]. In some situations, the BIC continues to increase as the number of groups increases [9], when the additional groups do not represent clinically distinct, valid subgroups. Different guidelines are often employed, such as no group should consist of less than 5% of the sample, the average group posterior probability has to be at least 0.70, the groups have to make “clinical sense”, and subject-specific judgments [2, 31].

2.2 METHODS OF CROSS-VALIDATION (CV)

2.2.1 K-Fold CV

K-fold CV is the basic form of CV, where other types are special forms of k-fold CV [7]. In k-fold CV the data is partitioned into k equally sized sets and k-iterations of training and validation are performed, where during each iteration a different set is excluded from modeling and the rest of the sets are combined to construct the model. The set that was excluded is then used to validate the constructed model. The most widely used versions of k-fold CV are five-fold and ten-fold. In order to fully understand k-fold CV, five-fold CV will be explained in more detail.

In five-fold CV the data is split into five roughly equal sets (Figure 1). The first step is to exclude one set and construct a model by combining the four other sets. In Figure 1, we first exclude the fifth set and combine sets one, two, three, and four to construct a model. Once the best-fit model is constructed for these four sets, set five is used to validate the constructed model. The information from the constructed model, will be used to predict the outcome of subjects in set five. Because the true values for these subjects’ outcome are actually known, an estimate of how well the model is performing on the fifth set can be determined by calculating the absolute deviation between the actual values and the predicted values:

$$AD_{5,i} = \left| y_i - \hat{y}_i^{[-5]} \right|.$$

Here $AD_{5,i}$ is the estimated error of the i^{th} subject who was randomly assigned to the fifth set, where $i = 1, \dots, n_5$. The actual value for the i^{th} subject's outcome is denoted as y_i and the predicted value of their outcome from the model that was constructed while excluding the fifth set is $\hat{y}_i^{[-5]}$. The absolute deviation is calculated for every subject in the fifth set and a mean absolute deviation is found by averaging across all of the subjects in the fifth set:

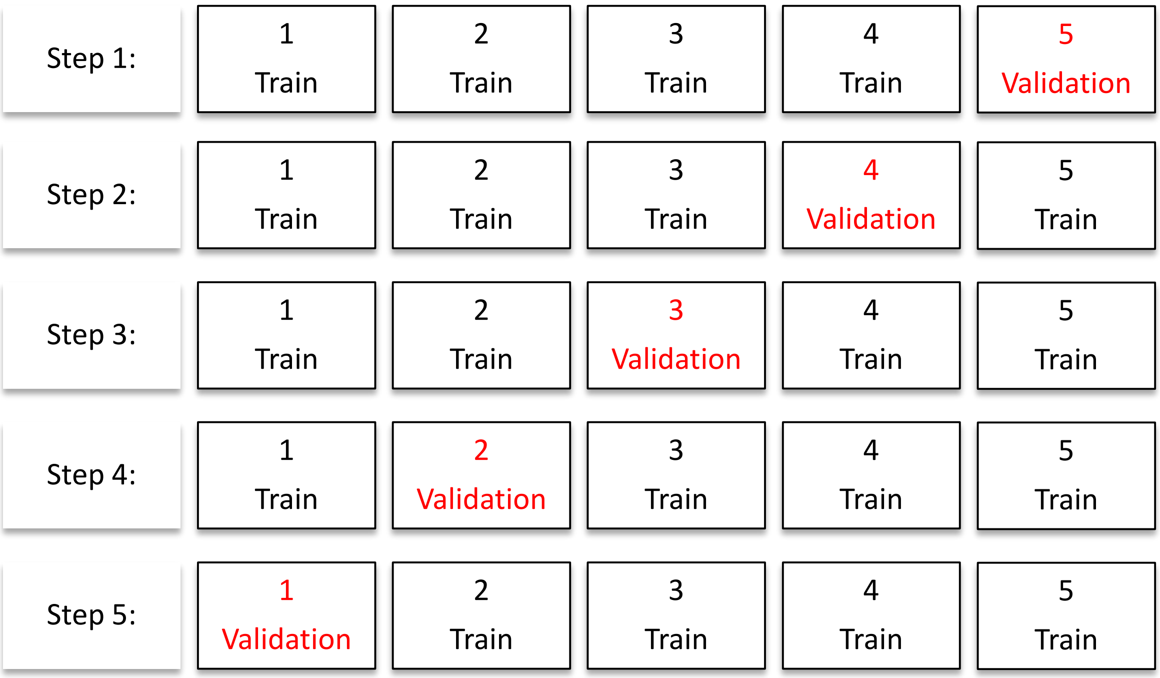
$$AD_5 = \frac{1}{n_5} \sum_{i=1}^{n_5} AD_{5,i}.$$

This gives the user an idea of how well their training model predicts future unseen data. This procedure is repeated four more times, each time excluding a different set of the data to use for validation and using the remaining 80% of the data for the training model (Figure 1). Once all calculations are completed for every set, an overall mean absolute deviation can be calculated by taking the grand mean of the average absolute deviations from each set of the data:

$$AD = \frac{1}{5} \sum_{w=1}^{5} AD_w,$$

where AD_w denotes the absolute deviation from the w^{th} set ($w = 1, \dots, n_w$).

There exist others ways of measuring the model fit in CV for continuous data including, median absolute deviation, mean squared error, paired t-test, and more [8, 32]. This paper focuses on mean absolute deviation.



Step 1:	1 Train	2 Train	3 Train	4 Train	5 Validation
Step 2:	1 Train	2 Train	3 Train	4 Validation	5 Train
Step 3:	1 Train	2 Train	3 Validation	4 Train	5 Train
Step 4:	1 Train	2 Validation	3 Train	4 Train	5 Train
Step 5:	1 Validation	2 Train	3 Train	4 Train	5 Train

Figure 1. Illustration of five-fold Cross-validation Process

2.2.2 Repeated K-fold Cross-validation

Repeated k-fold CV is the process of performing k-fold CV multiple times, where each time the data is reshuffled and re-stratified before each round. A commonly used version of repeated k-fold CV is five by two-fold CV [7]. In five by two-fold CV, available data is split up into two non-overlapping sets. The first set is held out and the second set is used to construct a model. Once the best-fit model is constructed, the first set is used to validate the constructed model. The AD between actual and predicted values of the outcome is calculated and this process is repeated, by constructing a model with the first set and validating the constructed model with the second set instead. Once this process is complete, the data is randomly split into two different groups and the process of training and validating is repeated on the two newly stratified groups. This is repeated three more times, each time randomly splitting the data up into two new groups.

The process of five by two-fold CV differs from five-fold CV in two ways: (1) the data is split into two groups instead of five and (2) once the process of two-fold CV is performed the data is re-stratified into two different groups and the process is repeated, so that there are five iterations of two-fold CV. An advantage of repeated k-fold CV is that the number of estimates has increased, smoothing the estimates, but a disadvantage is that an unreliably low estimate for the variance is given. Another disadvantage is that the training and testing data are now overlapping between each round of k-fold CV, as well as the training data is overlapped within each round of k-fold CV [7].

2.2.3 Leave one out Cross-validation (LOOCV)

LOOCV (also known as n-fold CV or Jackknife resampling) is a special case of k-fold CV, where k is the number of subjects in the available data. The process of LOOCV is first excluding one subject from the dataset, and then constructing a model using the remaining N-1 subjects. Once the best-fit model is constructed, the model is tested on the single subject that was excluded. This process is repeated for every subject in the data set. An advantage of this method of CV is that it results in unbiased estimates of the generalizability of a model's performance and disadvantages are it results in a very large variance of the performance estimate, since the estimate was based on a single subject and the process is computationally intensive for large datasets [7].

2.2.4 How CV is implemented in GBLTM for a censored normal model

SAS Proc TRAJ easily enables the addition of CV to GBLTM. An example of how this is done will be illustrated again using five-fold CV. Once the data is split into five roughly equal sets, the first step (Figure 1) is to combine sets one, two, three, and four to construct a training model, while setting set five aside for validation. Proc TRAJ will be used to construct the best-fit training model for the available data in sets one through four. Once the best-fit model is found, the MLE for each parameter in the model is easily obtained and outputted to a SAS dataset using the "OE" option in the proc statement. The MLEs are used to calculate the predicted values of the outcome at every measured time point for the subjects in the fifth set (the set excluded during model building process).

The MLE for the intercept and slope of each group ($\hat{\beta}_{0k}$ and $\hat{\beta}_{1k}$, respectively) are used to calculate $\hat{\mu}_{ijk}$, the MLE of the expected value of the repeated measures outcome, Y_{ijk}:

$$\hat{\mu}_{ijk} = \hat{\beta}_{0k} + \hat{\beta}_{1k} time_{ij} + \hat{\beta}_{2k} time_{ij}^2.$$

The MLE of $\theta = (\theta_1, \dots, \theta_k)$ and $\gamma = (\gamma_1, \dots, \gamma_k)$, denoted as $\hat{\theta} = \left(\hat{\theta}_1, \dots, \hat{\theta}_k\right)$ and $\hat{\gamma} = (\hat{\gamma}_1, \dots, \hat{\gamma}_k)$, are also available in the SAS dataset using the "OE" option. These MLEs are used to find an estimate of the posterior probability of group membership:

$$\hat{\pi}_i^k = \hat{P}r(C_i = k | Z_i = z_i) = \frac{exp\{\hat{\theta}_k + \hat{\gamma}_k' z_i\}}{\sum_{l=1}^K exp\{\hat{\theta}_l + \hat{\gamma}_l' z_i\}},$$

Proc TRAJ does not provide the MLE for θ_1 and γ_1 because these values are always zero. Therefore, the estimate for the posterior probability of being in group one simplifies to:

$$\hat{\pi}_i^1 = \hat{P}r(C_i = 1 | Z_i = z_i) = \frac{1}{1 + \sum_{l=2}^K exp\{\hat{\theta}_l + \hat{\gamma}_l' z_i\}},$$

and an estimate of the posterior probability for all other latent subgroups is calculated by :

$$\hat{\pi}_i^k = \hat{P}r(C_i = k|Z_i = z_i) = \frac{exp\{\hat{\theta}_k + \hat{\gamma}_k' z_i\}}{1 + \sum_{l=2}^{K} exp\{\hat{\theta}_l + \hat{\gamma}_l' z_i\}},$$

where $k = 2, \dots, K$.

Once the MLE for the posterior probability of group membership and the MLE of the expected value of Y_{ijk} are calculated, an estimate of the mean of Y_{ij} can be calculated for each subject as a weighted average:

$$\hat{y}_{ij}^{[-w]} = \sum_{k=1}^{K} \hat{\pi}_i^k \hat{\mu}_{ijk}.$$

The estimate of the mean of the repeated measures outcome, Y_{ij}, also known as the predicted values, are then compared to the actual measurements of Y_{ij} for each subject in the fifth set. In this thesis we compare the predicted values to the observed values using absolute deviation (AD):

$$AD_{wij} = \left| y_{ij} - \hat{y}_{ij}^{[-w]} \right|.$$

Once the absolute deviation is calculated at every time point for each subject the average across all time points for each individual i is calculated:

$$AD_{wi} = \frac{1}{J_i} \sum_{j=1}^{J_i} AD_{wij}.$$

Lastly, the grand mean AD across all subjects is calculated:

$$AD_w = \frac{1}{N} \sum_{i=1}^{N} AD_{wi}$$

This process is then repeated for each step of five-fold CV (Figure 1), where at each iteration a new set is held out for validation, resulting in five values of AD. The average AD across each set is calculated as:

$$AD = \frac{1}{5}\sum_{w=1}^{5} AD_w.$$

Once five-fold CV is completed, the whole process is repeated, where the only thing that differs in the models is the number of latent subgroups. The model with the number of latent subgroups that results in the smallest AD should be chosen.

3.0 CLINICAL APPLICATION

Prior research in patients with hepatitis C (HCV) found that those who reported poor sleep quality preceding the start of treatment were at a higher risk for developing major depressive disorder (MDD) than those with good sleep quality [33]. Whether this association holds generally among the entire study population or whether it was driven by a specific subgroup is currently unknown. Therefore, GBLTM was applied to the HCV study in order to address the following two aims: first, to determine whether patterns of change in depressive symptoms formed clinically meaningful, distinct subgroups, and second, to test the hypothesis that subjectively assessed sleep disturbances are associated with less favorable depression trajectories.

3.1 BACKGROUND OF HEPATITIS C (HCV) STUDY

3.1.1 Recruitment

Patients with HCV were recruited at a University of Pittsburgh Medical Center that specializes in liver diseases prior to starting treatment of pegylated interferon-alpha2 (PEG-IFN-α2a: 135 μg/week or PEG-IFN-α2b: 120 or 150 μg/week) and oral ribavirin. Those who were recommended for interferon-α (IFN-α) treatment by their hepatologist were asked to participate in a research study where they would be followed for up to four months in order to investigate the association of INF-α and depression. All subjects provided written informed consent as per protocol by the University of Pittsburgh Institutional Review Board.

3.1.2 Exclusion Criteria

Subjects were excluded if they had an active Axis-I disorder (e.g. mood, anxiety, psychotic, impulse control, or drug/alcohol use disorders) within six months prior to starting treatment because the primary interest of the study was to detect interferon-induced depression (IFN-MDD). An Axis-I disorder was assessed using the Structural Clinical Interview for DSM-IV Axis I Disorders (SCID-I) [34]. Subjects were also excluded if they had a history of prior depressive episodes where subjects had been in remission for less than six months, known neurological disease, known active inflammatory disorders other than HCV, or if they were taking medications known to influence the immune system (e.g. corticosteroids, anticonvulsants, and/or antipsychotics).

3.1.3 Patient Characteristics

There were 124 patients diagnosed with HCV who began IFN-α therapy. Patients were mostly male (64.5% (n=80)) and Caucasian (90.3% (n=112)) with a mean (standard deviation) age of 45.7 (12.3) years. Other than HCV, subjects had medical burden that was comparable to those in their 40s without HCV, based on the Cumulative Illness Rating Scale-Geriatric [35]. Subjects who developed major mood disorder during treatment were given a psychiatric intervention, consisting of an antidepressant or mood stabilizing medication, and their subsequent data was censored.

3.1.4 Variables of Interest

The amount of depressive symptoms a subject had was measured prior to starting treatment and at weeks 2, 4, 8, 12, and 16 during treatment using the Beck Depression Inventory-II (BDI) [36]. The BDI consists of 21 questions where an individual can score 0-3 on each, resulting in a total range of 0-63. Scores on the BDI are categorized as minimal depression (0-13), mild depression (14-19), moderate depression (20-28), and severe depression (29-63). Subjects who scored a 15 or higher on the BDI during the course of treatment were administered an abbreviated SCID-I in order to detect a DSM-IV mood disorder.

Self-reported sleep quality was assessed prior to starting treatment using the Pittsburgh Sleep Quality Index (PSQI) [37]. The PSQI ranges from 0-21, where a higher score means worse sleep quality. A score below five on the PSQI typically indicates no problem with sleep and scores of five or higher indicate some sleep problems, where the median PSQI score of subjects diagnosed with insomnia is ten [37-39]. In order to account for sleep as a potential risk factor, BDI minus the sleep item (BDI-s) was used in analyses.

4.0 MODEL SELECTION USING BIC

4.1 DETERMINING THE NUMBER OF SUBGROUPS

The model selection process used in GBLTM is already well established using the Bayesian Information Criterion (BIC) [19, 23]. Examining the overall trajectories of depressive symptoms during the course of treatment it is clear that the HCV sample does not follow one single pattern (Figure 2). The number of latent subgroups of depressive symptoms over time was determined by examining the BIC. Five different models were considered, where each model consisted of a different number of latent subgroups; up to six latent subgroups were considered. A model with more than six groups was not considered because these models failed to converge. Each of the five models were examined by using first-degree polynomials and then again by using second-degree polynomials. A first-degree polynomial indicates that the trajectory is assumed to be a straight line. Similarly, a second-degree polynomial means that the trajectory is assumed to be a quadratic curve. When determining the optimal number of distinct subgroups in model selection, it is suggested that setting a second-degree polynomial for all trajectories is best because a quadratic form is very flexible in its ability to capture different rates of change [23]. Although, for completeness and because most of our data follows a linear pattern, a first-degree polynomial was examined as well.

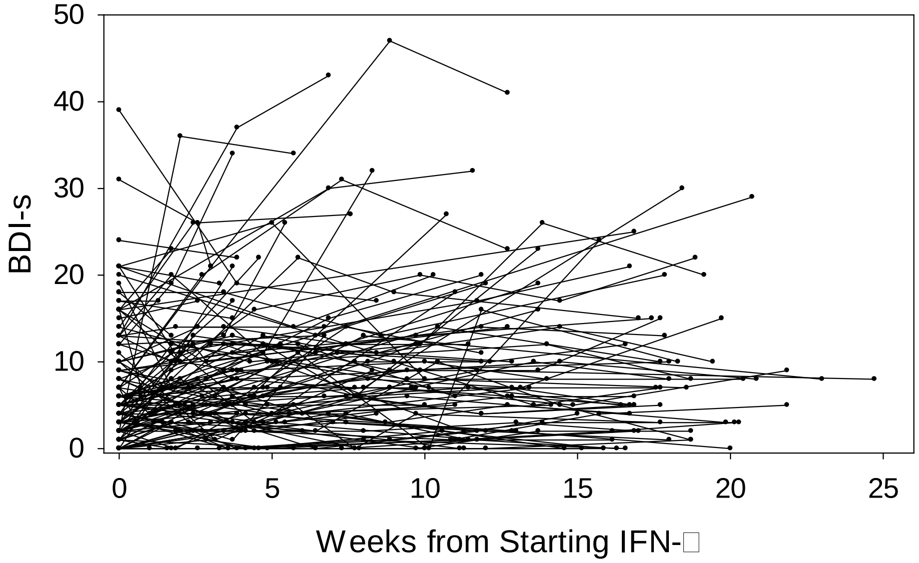


Figure 2. Individual trajectories of BDI-s prior to starting treatment and during treatment

In section 2.1.1, it is explained that the largest BIC is best, as long as it is a meaningful difference in BIC between the models that are being compared. Figure 3 displays the BIC for the five models with two to six subgroups when applying all linear terms and when applying all quadratic terms. The BIC increases monotonically until it reaches a six-group model, with more substantive increases occurring between groups two and three. After group three the slope of change in BIC decreases, but it is still apparent that the model is doing better up to a five-group solution. When comparing two models, the null model (or the simpler model) is the model with the smaller number of subgroups. Thus, when comparing a three-group to a two-group model, the null model would be the two-group model, and the log Bayes factor approximation is a way to judge the amount of evidence against the simpler model. In Table 2, each model was compared to the model that had one less subgroup. According to the criteria for the log Bayes factor approximation (Table 1) there exists "very strong" evidence against every null model except when comparing a six-group model to a five-group model. The six-group model failed to converge when all trajectories were set to a quadratic curve, furthering the evidence against a six-group model. Thus, it appears that five groups is the optimal number of latent subgroups to describe the HCV sample.

Table 2. Determining the number of subgroups using BIC

Number of Latent Subgroups	Number of groups in the Null Model	All Linear Terms		All Quadratic Terms	
		BIC	2(ΔBIC)	BIC	2(ΔBIC)
2		-1526.2		-1526.1	
3	2	-1489.5	73.4	-1491.8	68.6
4	3	-1482.1	14.8	-1485.3	13.2
5	4	-1473.5	17.3	-1474.7	21.2
6	5	-1478.8	-10.6	------	------

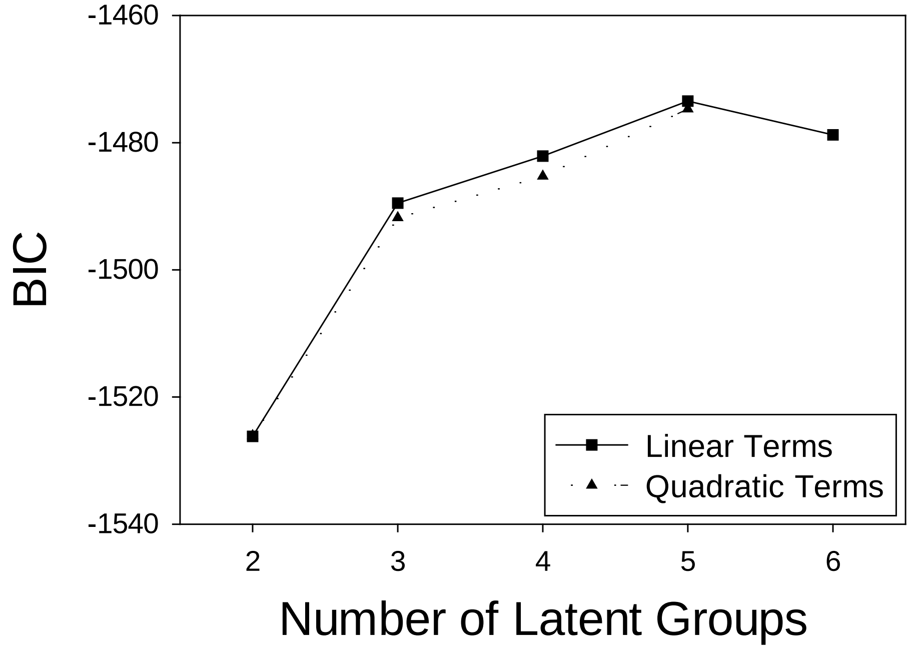


Figure 3. BIC for models with two through five subgroups

4.2 DETERMINING THE BEST POLYNOMIAL DEGREE FOR EACH TRAJECTORY

The next step is to determine the polynomial degree for each trajectory in the five-group model. This is done by examining every permutation of linear and quadratic terms, resulting in thirty-two five-group models examined (Table 3). For simplicity, BIC in Table 3 was sorted from best to worst, and then labeled model one to thirty-two. The difference between BIC for each model is illustrated in Figure 4, where it is apparent that model one is consistently the best. Whether or not the decline in BIC after model one is significant was again examined using the log Bayes factor approximation.

Every model was compared to model one and because model one contains mostly linear terms, it is the null model for every comparison except with model sixteen, which has all linear terms. The log Bayes factor approximation is negative for every comparison (except for model sixteen) indicating that there is no evidence against the null model: model one. When comparing model one to sixteen, there is “very strong” evidence against the null model: model sixteen. Therefore, model one was chosen as the best combination of polynomial degrees for a five-group model.

Table 3. Determining the best polynomial degree for each trajectory in a five-group model using BIC

Model	**Polynomial degree**	**BIC**	**$2\log_e(B_{10})$ $\approx 2*(\Delta \mathrm{BIC})$**
1	1 1 1 1 2	-1466.4	
2	1 2 1 1 2	-1467.8	-2.8
3	1 1 2 1 2	-1468.6	-4.2
4	1 1 1 2 2	-1468.8	-4.7
5	2 1 1 1 2	-1468.9	-4.8
6	1 2 2 1 2	-1470.0	-7.0
7	1 2 1 2 2	-1470.2	-7.5
8	2 2 1 1 2	-1470.3	-7.6
9	1 1 2 2 2	-1470.9	-8.9
10	2 1 2 1 2	-1471.0	-9.0
11	2 1 1 2 2	-1471.2	-9.5
12	1 2 2 2 2	-1472.3	-11.7
13	2 2 2 1 2	-1472.4	-11.8
14	2 2 1 2 2	-1472.6	-12.3
15	2 1 2 2 2	-1473.3	-13.7
16	1 1 1 1 1	-1473.5	14.0
17	2 2 2 2 2	-1474.7	-16.5
18	1 1 2 1 1	-1474.8	-16.7
19	1 2 1 1 1	-1474.8	-16.7
20	1 1 1 2 1	-1475.8	-18.7
21	2 1 1 1 1	-1475.9	-18.8
22	1 2 2 1 1	-1476.4	-20.0
23	1 2 1 2 1	-1477.1	-21.4
24	2 2 1 1 1	-1477.2	-21.6
25	1 1 2 2 1	-1477.4	-21.8
26	2 1 2 1 1	-1477.5	-22.1
27	2 1 1 2 1	-1478.2	-23.5
28	1 2 2 2 1	-1478.7	-24.5
29	2 2 2 1 1	-1478.9	-24.8
30	2 2 1 2 1	-1479.5	-26.2
31	2 1 2 2 1	-1479.8	-26.7
32	2 2 2 2 1	-1481.1	-29.3

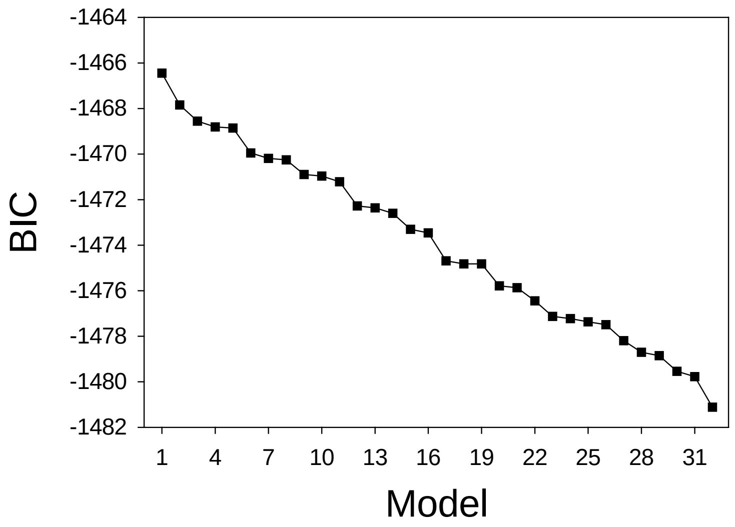


Figure 4. Determining the best polynomial degree for each trajectory in a five-group model using BIC

4.3 ADDING RISK FACTORS TO THE BEST-FIT MODEL

After the optimal number of groups and the best-fit polynomial degree for each group is determined, risk factors can be added. PSQI was added to the best-fit five-group model as a potential risk factor in order to answer the hypothesis of interest. When adding risk factors in Proc TRAJ it is advised to use well-informed start values [17], which help the program locate the true MLEs. Start values are the MLE of the intercept and slope for each trajectory group, the MLE of the variance, and the MLE of theta in the posterior probability. All start values are obtained from running the model with no risk factors. The program outputs the values to the log, which can then copied and pasted in the Proc TRAJ code with risk factors using the “start” statement. A MLE for the risk factors also have to be specified, so if you have a five-group model four start values have to be listed for each risk factor added. However, the start values for the risk factors are not outputted in the log, so it is suggested to use zero for their values.

When adding in the start values in addition to PSQI as a potential risk factor to the best-fit five-group model, the model runs with no convergence issues. The average and predicted group trajectories are illustrated in Figure 5. Groups one and two are very similar, both groups on average have very low depressive symptoms prior to starting treatment and these low symptoms remain, on average, constant during the course of treatment. Together, these two groups make up about 38% of the HCV sample. Group three makes up about 30% of the sample, and on average also experiences consistently low depressive symptoms that are slightly higher than groups one and two, but are still relatively and persistently mild during the course of treatment. The fourth group is made up of about 24% of the sample; on average, they experienced minimal depression prior to starting treatment, but steadily increase to moderate depression by four months of IFN-α. The fifth group consists of a small subgroup of the sample, about 8%. They experienced mild depression on average prior to starting treatment and rapidly increase to severe depression just after one month on IFN-α. All subjects in the fifth group were diagnosed with a mood disorder and were excluded from the study by three months of IFN-α because they needed antidepressant treatment.

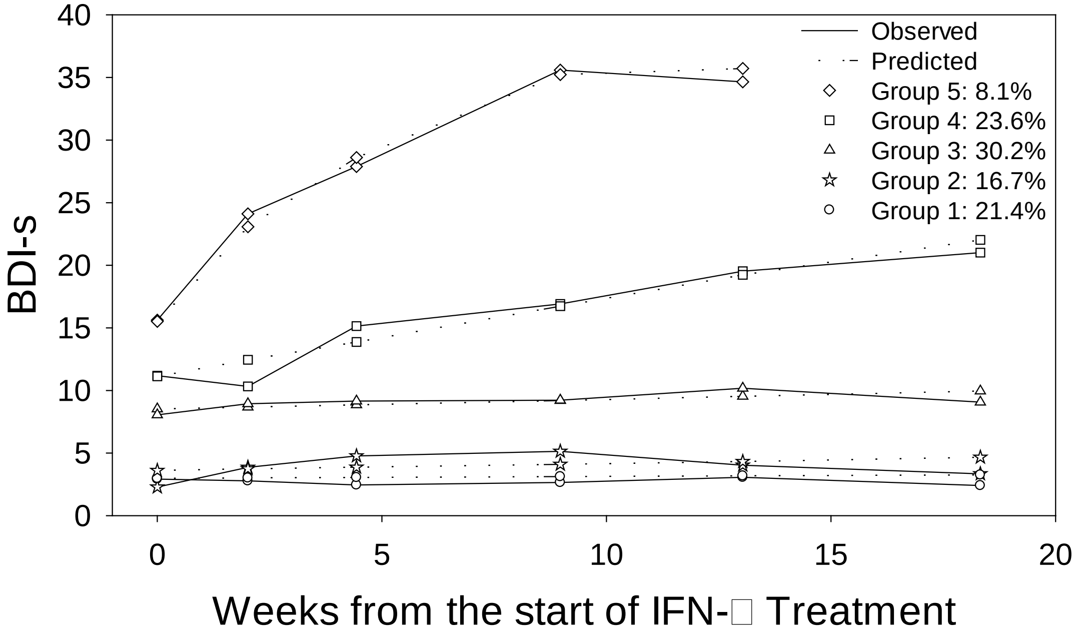


Figure 5. Best-fit five-group model of BDI-s prior to starting treatment and during treatment when including PSQI at a baseline risk factor

4.4 BIC SUGGESTING TOO MANY GROUPS

When considering the average group intercept and slope of change in Figure 5 it appears that there exist a smaller number of distinct latent subgroups in the HCV sample. Investigating the individual trajectories by each of the five groups (Figure 6), it appears that groups one, two, and some individuals from group three could be combined to one group. Each individual's trajectory in group one and two do not differ by much. Some subjects in group three also appear to have a consistently low trajectory as well. Group four appears to consist of a few subjects that could be combined to group five, and the majority of group four could be combined with some subjects in group three, to get a better representation of a group that has a steady increase in their depressive symptoms over the course of treatment. These observations are consistent with other studies [1-

6], which suggests BIC indicates a larger number of groups than what is needed to describe the data in a meaningful and parsimonious fashion.

When PSQI is added as a baseline risk factor along with the start values from the best-fit model with no risk factors included, the observed and predicted trajectories differ (Figure 5 compared with Figure 14 in Appendix A). Groups two, three, and four are altered with addition of a baseline risk factor, the slope of one of the trajectory groups even changes direction. The addition of a baseline risk factor should not alter a model this immensely; therefore furthering the evidence against the five-group model.

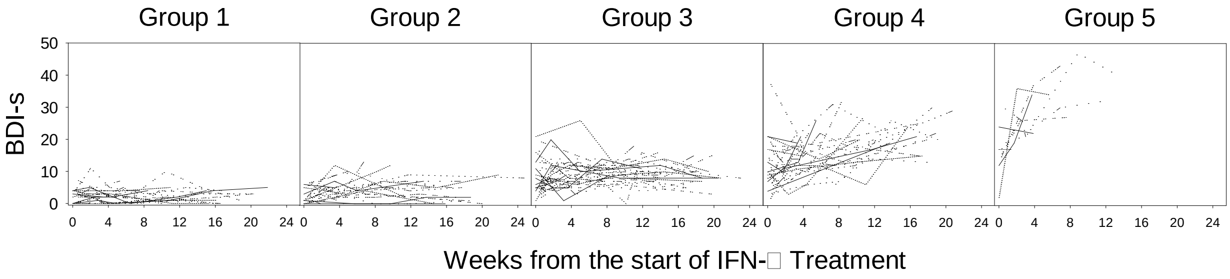


Figure 6. Individual trajectories of BDI-s prior to starting treatment and during treatment by the best-fit five group model when including PSQI as a baseline risk factor

5.0 CROSS-VALIDATION RESULTS

The GBLTM analysis on the sample of 124 patients treated for HCV is validated using three types of CV: n-fold CV, five-fold CV, and five by two-fold CV. Table 4 displays the overall mean absolute deviation (AD) between observed and predicted values of BDI-s for two through five latent subgroups for each of these types of CV. Six latent groups were not included, due to convergence issues reported in section 4.1. The results are split up into two sections: setting all trajectories to a linear curve and setting all trajectories to a quadratic curve.

CV was consistently a more conservative approach to determining the optimal number of latent subgroups (compared with BIC). All three types of CV suggest that a two-group model results in the smallest error (Table 4, Figure 7). The AD steadily increases as the number of latent subgroups increases. Although, the increase in AD is very small: there is a 9% relative increase in AD for a three-group model compared to a two-group model, a 6% relative increase in AD for a four-group model compared to a three-group model, and a 4% relative increase in AD for a five-group model compared to a four-group model.

Table 4. Absolute Deviation between observed and predicted values of BDI-s for different number of latent subgroups using cross-validation

Polynomial Degree for each group	Number of Latent Classes	AD		
		n-fold CV	5-fold CV	5x2-fold CV
First-degree	2	6.603	6.583	6.592
	3	7.206	7.202	7.210
	4	7.639	7.640	7.640
	5	7.950	7.943	7.965
Second-degree	2	6.603	6.583	6.601
	3	7.232	7.238	7.241
	4	7.692	7.693	7.688
	5	7.988	7.994	8.159

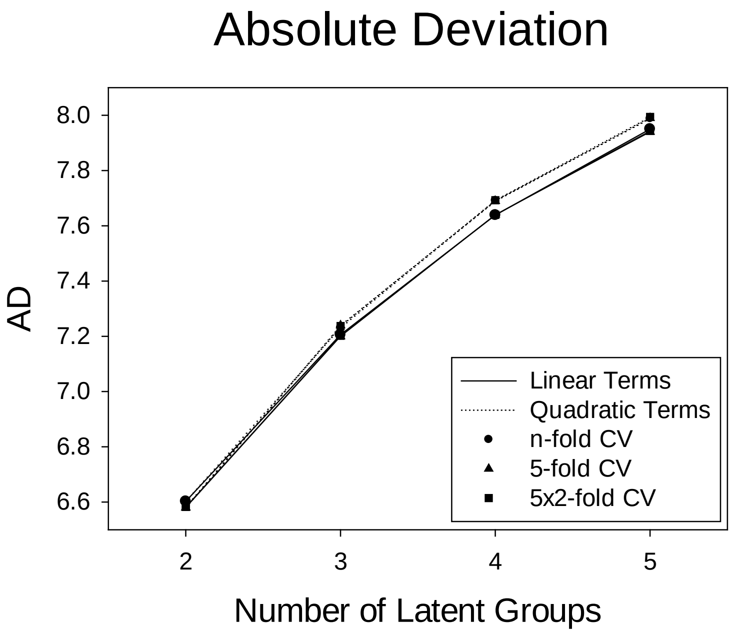


Figure 7. Absolute Deviation between observed and predicted values of BDI-s for different number of latent subgroups using cross-validation

5.1 INVESTIGATING THE TRUE NUMBER OF LATENT SUBGROUPS

CV suggests a two-group model, whereas BIC suggests a five-group model, but when visually inspecting the data, a three-group model appears to best illustrate the distinct subgroups in the sample. A four-group model is very similar to the three-group model, adding in what appears to be a redundant non-distinct group (Figure 8 and 9 compared with Figure 10 and 11). A portion of the individuals that make up the second group of the four-group model could be combined with the first group and the others that have a larger increasing slope could be combined with the third group. A few individuals with a large increase in depressive symptoms in the third group could be combined to the fourth group. Thus, concluding a three-group model appears better than a four-group model.

When examining a two-group model, there is a group of subjects that experience low average depressive symptom prior to starting treatment that has a slight overall increase during treatment and another group that starts off with a higher overall average depressive symptom count, that is steadily increasing at a slightly higher rate of change compared with the first group, but decreases around the last time point (Figure 12). The decrease in the trajectory of the second group is occurring because the individuals who are rapidly becoming depressed are excluded before the end of the study (Figure 13). This can cause a misinterpretation of the second group because it appears that overall the subjects are eventually becoming less depressed around 16 weeks of treatment, when in fact the opposite is true. The second group is also very heterogeneous, where there is not one overall pattern of the group (Figure 13), similar to what was found when examining the overall plot of individual trajectories of depressive symptoms for every subject (Figure 2). Thus, it still remains apparent that the three-group model appears to be best capturing all information in the HCV sample in the most parsimonious way.

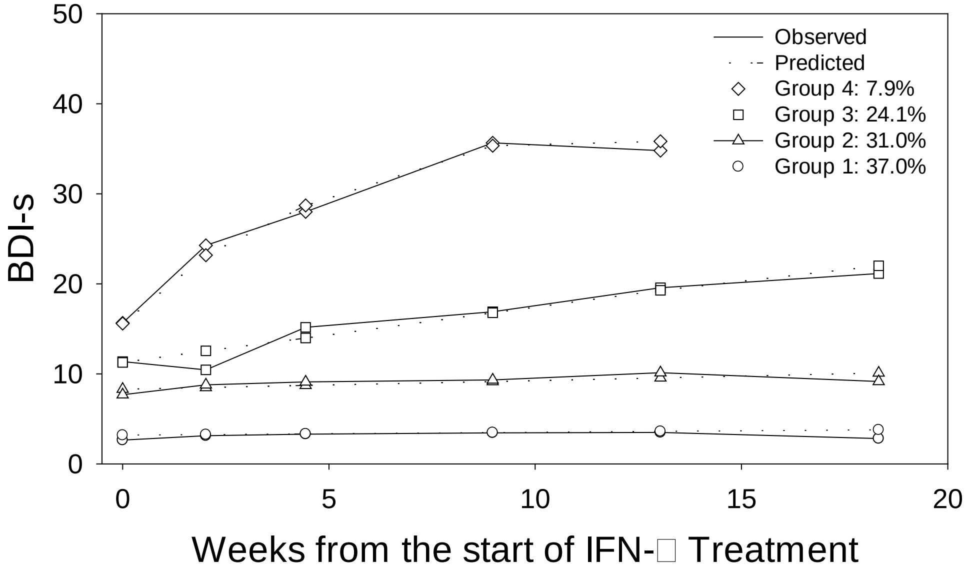


Figure 8. Best-fit four-group model of BDI-s prior to starting treatment and during treatment when including PSQI as a baseline risk factor

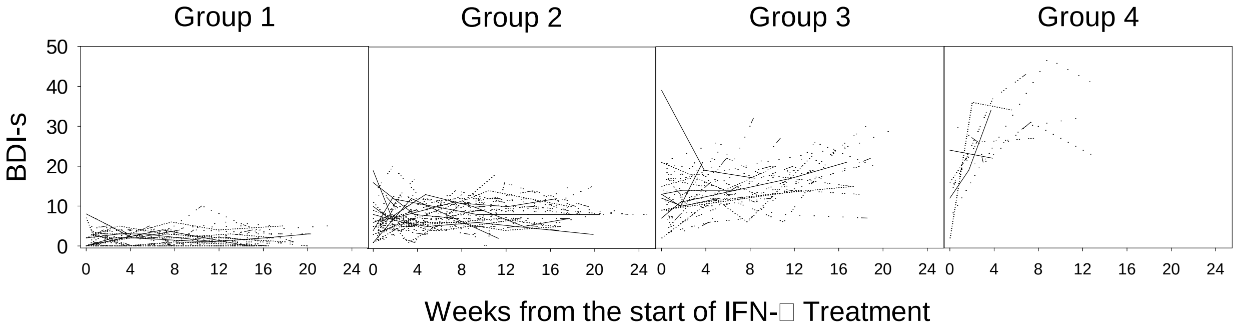


Figure 9. Individual trajectories of BDI-s prior to starting treatment and during treatment by the best-fit four-group model when including PSQI as a baseline risk factor

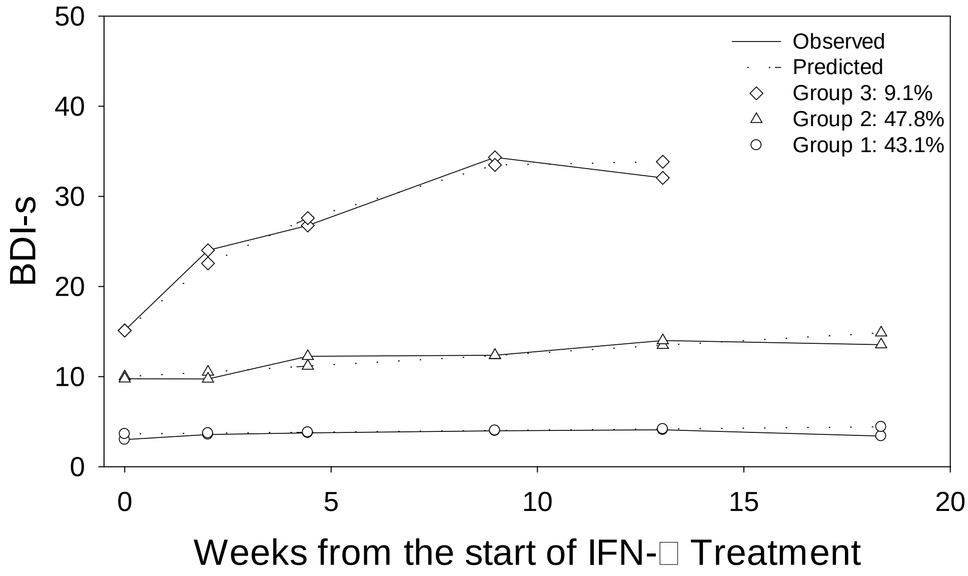


Figure 10. Best-fit three-group model of BDI-s prior to starting treatment and during treatment when including PSQI at a baseline risk factor

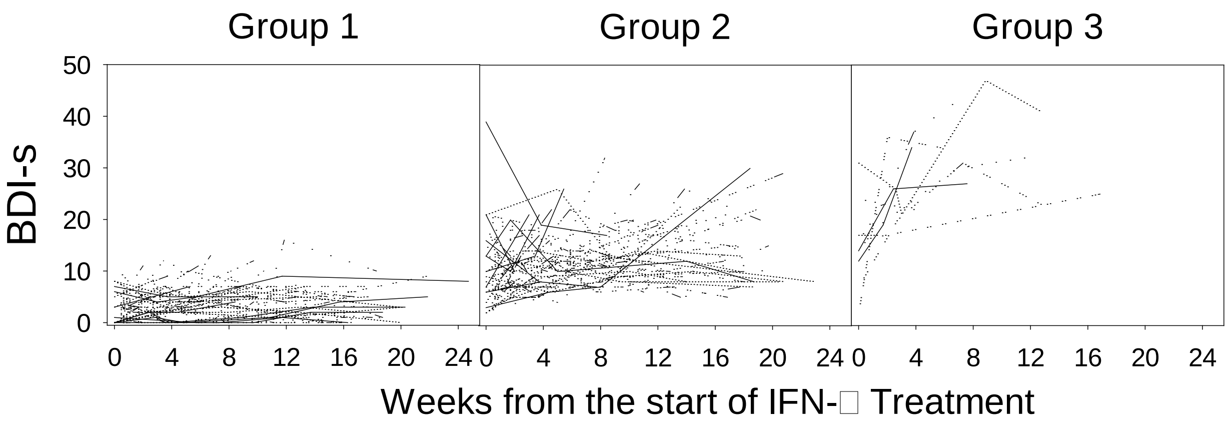


Figure 11. Individual trajectories of BDI-s prior to starting treatment and during treatment by the best-fit three-group model when including PSQI as a baseline risk factor

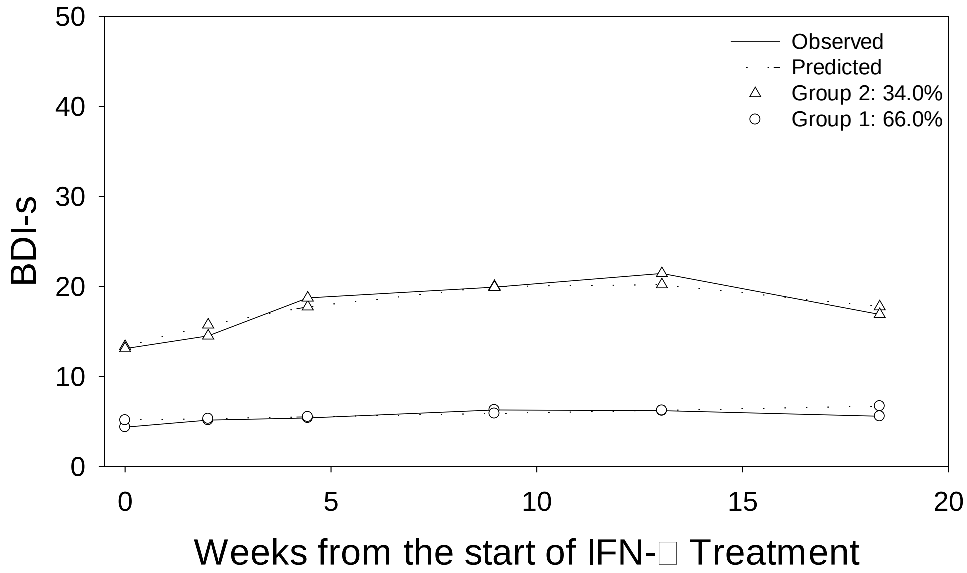


Figure 12. Best-fit two-group model of BDI-s prior to starting treatment and during treatment when including PSQI at a baseline risk factor

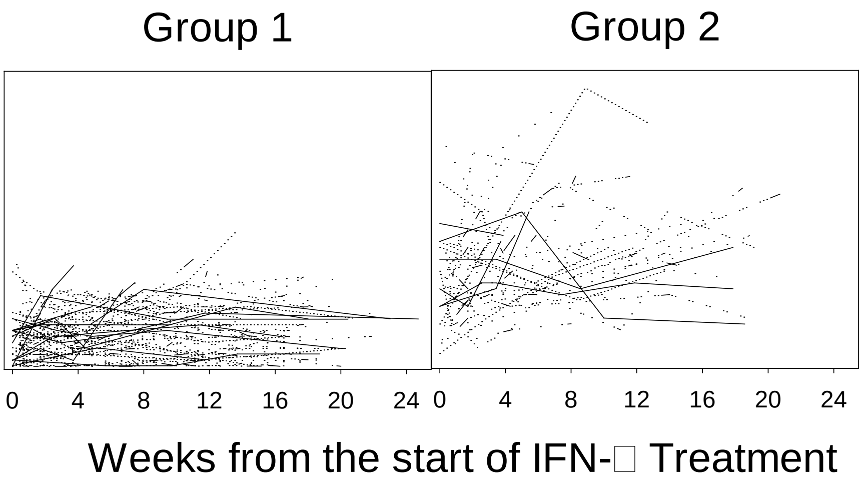


Figure 13. Individual trajectories of BDI-s prior to starting treatment and during treatment by the best-fit two-group model when including PSQI as a baseline risk factor

6.0 RESULTS OF A THREE-GROUP MODEL

In the best-fit three-group model, the three groups were labeled as group 1: “non-depressed”, group 2: “slow increase”, and group 3: “rapid increase” (Figure 10). The non-depressed and slow increase trajectories were best approximated as a linear increase, whereas a quadratic increase was best for the rapid increase trajectory. The non-depressed group contained 43.1% (N=54) of the sample who started out with a low overall group mean BDI-s of 3.01 at baseline and experienced very little change over the course of treatment (Table 5). The slow increasers were made up of 47.8% (N=59) of the sample and started out with an overall group mean BDI-s of 9.77 which steadily increased throughout the course of treatment to mild depression by week 12 of IFN-α treatment (14.01). The rapid increasers consisted of 9.1% (N=11) of subjects, with a mean BDI-s at baseline categorized as mild depression (15.13), which reached severe depression (34.23), on average, after 8 weeks of IFN-α treatment. All rapid increasers were excluded from the study by 16 weeks of IFN-α because they began antidepressant treatment.

Table 5. Mean and confidence intervals of BDI-s over time by subgroups while adjusting for baseline PSQI

		Non-depressed N=54 Mean (95% CI)	Slow Increase N=59 Mean (95% CI)	Rapid Increase N=11 Mean (95% CI)
Pre IFN-α	Baseline	3.01 (2.4, 4.9)	9.77 (8.5, 11.5)	15.13 (11.9, 18.3)
During IFN-α Treatment	Week 2	3.57 (2.6, 4.9)	9.74 (9.2, 11.9)	24.02 (20.2, 24.9)
	Month 1	3.76 (2.7, 4.9)	12.26 (9.9, 12.4)	26.77 (24.8, 30.4)
	Month 2	3.98 (2.8, 5.2)	12.38 (11.1, 13.6)	34.34 (28.6, 38.4)
	Month 3	4.09 (2.7, 5.7)	14.01 (11.9, 15.1)	32.06 (29.3, 38.4)
	Month 4	3.40 (2.3, 6.6)	13.55 (12.8, 17.0)	------

Age and sex were not associated with group membership; therefore excluding them from the three-group model reported did not drastically alter the results. PSQI significantly predicted group membership (Table 6), where worse sleep was associated with a less favorable trajectory. With one unit increase in PSQI, the odds of having a slow increase in depressive symptoms instead of no change (non-depressed group) increases by 31% and the odds of having a rapid increase in depressive symptoms compared to no change increases by 56%. With one unit increase in PSQI, the odds of having a rapid increase in depressive symptoms instead of a slow increase increases by 19%. The average posterior probability for each group in the model was greater than 0.9, thus concluding an adequate model fit.

Table 6. Baseline PSQI as a risk factor of group membership

Reference Group: Non-depressed			
Group	Variable	Estimate (SE), p-value	Odds Ratio
Slow Increase	Constant	-1.58 (0.51), p=0.002	----
	PSQI	0.272 (0.08), p<0.001	1.31
Rapid Increase	Constant	-4.91 (1.1), p<0.001	----
	PSQI	0.448 (0.11), p=0.001	1.56
Reference Group: Slow Increase			
Group	Variable	Estimate (SE)	Odds
Rapid Increase	Constant	-3.33 (1.0), p=0.001	----
	PSQI	0.176 (0.09), p=0.05	1.19

7.0 DISCUSSION

CV was applied to GBLTM for censored normal models as an alternative method to BIC for model selection. All three types of CV (n-fold CV, five-fold CV, and five by two-fold CV) consistently indicated a two-group solution resulted in the smallest AD between observed and predicted values, whereas BIC suggested that a five-group model was best. Therefore, when compared with BIC, CV is a more conservative approach to model selection in GBLTM for the HCV sample.

Model selection with BIC in GBLTM is a complex process with many iterations of re-fitting the model. It appears that there is no one metric to use in the model selection process. Visually examining the groups from selected best-fit models can provide judgment on whether the model (BIC or CV suggests) is actually fully capturing the distinct latent subgroups of the sample. Spaghetti plots are a very useful tool for depicting distinct groups in a sample. As shown above, once we plotted the individual trajectories for each subject by the best-fit five-group model, it was clear that the data actually followed a three-group model.

However, BIC and CV should always be considered when determining the optimal number of latent subgroups in a model. Both tools for model selection enable the user to make an educated judgment regarding the number of latent subgroups in their sample. CV may be more useful than BIC in situations where users may want the smallest number of latent subgroups possible to describe the heterogeneity in their data. In this case they would be best served with

CV as the model selection metric of choice. On the other hand, BIC may be more useful when clinically distinct subgroups are continuously identified.

With an increasing use of GBLTM, there is a greater need for more standardized, quantitative, and rigorous methods for selecting the number of latent subgroups in a population. This thesis contributes to public health research by exploring different statistical tools used for determining the optimal number of latent subgroups in GBLTM. We assessed CV as an alternative tool to BIC for model selection in GBLTM and found it is a conservative approach compared with BIC for our sample. CV may be more useful to researchers when they are interested in selecting the smallest number of groups to explain differences in their target patient population and time course.

7.1 LIMITATIONS

Whether or not CV is truly a more conservative approach to model selection compared with BIC is still unknown. This may be true for only this sample of HCV subjects. CV and BIC should be examined using other datasets, as well as using simulations. Another limitation of our study is that the five-group model becomes very unstable once a baseline risk factor is added, causing a change in the five trajectories. The addition of baseline risk factors should not significantly alter the trajectories, indicating that the five-group model without baseline risk factors might be displaying distinct subgroups of the HCV population, but needs to be determined using a larger sample of patients.

7.2 FUTURE WORK

In our data, CV is a conservative approach to model selection in GBLTM when compared with BIC. Although, future work using various simulations can shed light on the generalizability of our findings. One simulation should create data that consists of a various number of actual trajectory groups to examine whether BIC and/or CV correctly classifies the true groups. Another simulation should create data that consists of no distinct groups to examine how BIC and CV react.

With regard to CV specifically, a meaningful difference in AD between two models is unknown. A useful criterion to determine a meaningful difference in AD is needed to examine whether the different models are truly differing in their predictability. In also appears that the criterion for a meaningful difference in BIC between two models is very lenient. One possibility is to manipulate the equation for BIC so that there is a larger penalty term when used in GBLTM. An interpretation similar to scree plots could be useful to BIC as well. With scree plots, one chooses the first value where the measurements start to level off. For BIC, this would mean choosing the number of latent subgroups where the BIC starts to level off, even if the addition of a latent group represents a meaningful difference in the BIC according to the log Bayes factor approximation. Using this interpretation, a three-group model would be suggested for the HCV sample (Figure 3).

8.0 CONCLUSION

CV offers a conservative approach to model selection in GBLTM when compared with BIC. However, both BIC and CV should be used with caution during model selection in GBLTM. A visual inspection of the data should always be performed before concluding which model is the best-fit. Overall group mean observed and predicted values should be plotted and inspected as well as the overall and by group spaghetti plots. Examining these plots can give researchers an idea regarding whether or not the large number of latent subgroups suggested by BIC are actually distinct subgroups in the sample.

Examining solutions suggested by CV, along with the associated plots gives one insight on whether or not a small number of groups is picking up on all of the information shown in the data. Missing data should also be inspected as with every repeated measures analysis. When plotting overall group mean observed and predicted values, if a group that experiences a less favorable trajectory is suddenly on average doing better at the end of the study, carefully examine the data to see if this is simply due to drop out towards the end of the study. The subjects that are having a worse experience are more likely to drop out of the study, thus making the trajectory appear as if the group gets better towards the last few visits. BIC, CV, and an educated judgment are useful tools for model selection when determining the optimal number of latent subgroups in a sample using GBLTM.

APPENDIX A: BEST-FIT MODELS WITHOUT BASELINE RISK FACTORS

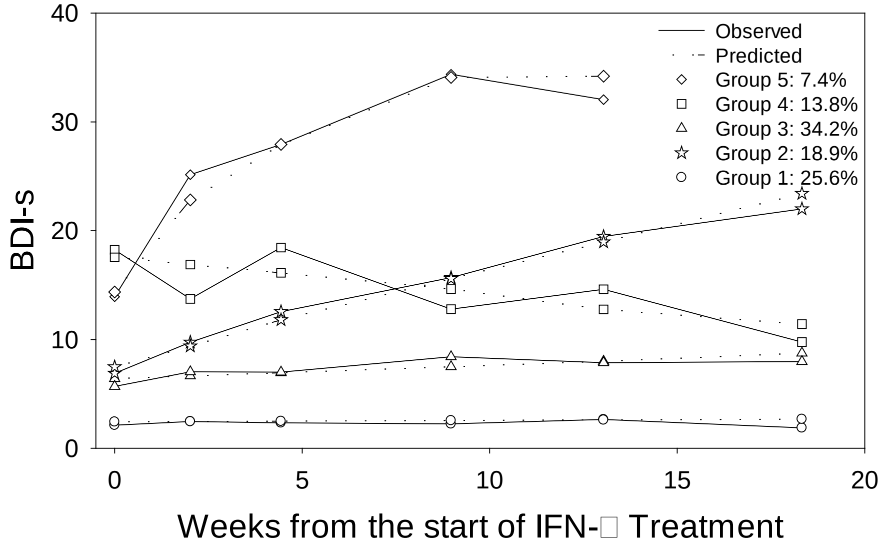


Figure 14. Best-fit five-group model without baseline risk factors

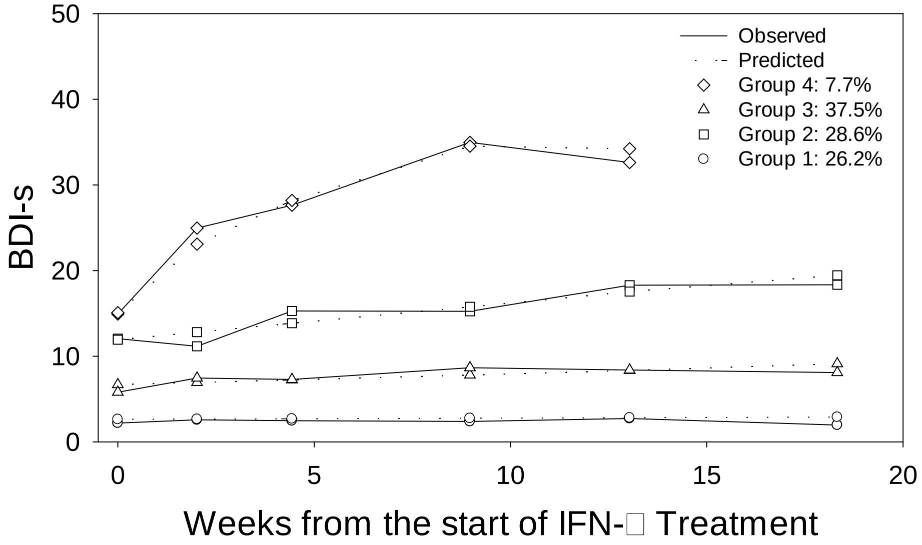


Figure 15. Best-fit four-group model without baseline risk factors

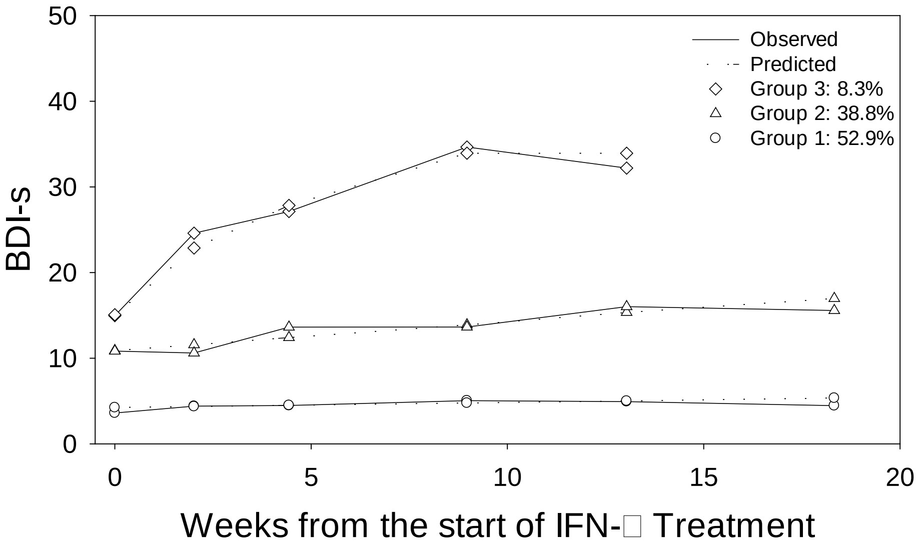


Figure 16. Best-fit three-group model without baseline risk factors

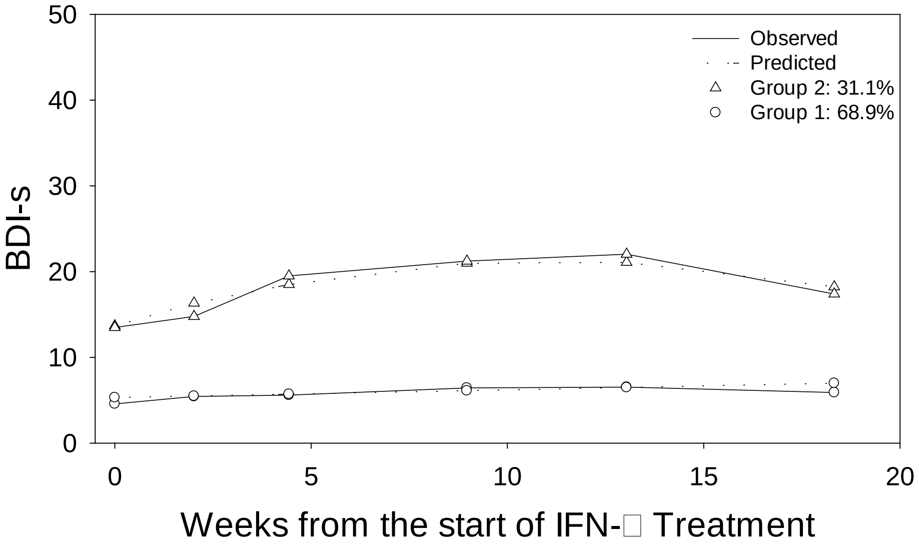


Figure 17. Best-fit two-group model without baseline risk factors

APPENDIX B: CROSS-VALIDATION SAS MACROS

B.1 N-FOLD CV

```
%macro nfold(datafile=, N=,ngroups=, order=, num_timepts=, y1=, y2=, y3=, y4=, y5=,
y6=,time1=,time2=,time3=,time4=,time5=,time6=,min=,max=);

%let outcome1=%upcase(&y1);
%let outcome2=%upcase(&y2);
%let outcome3=%upcase(&y3);
%let outcome4=%upcase(&y4);
%let outcome5=%upcase(&y5);
%let outcome6=%upcase(&y6);

proc sort data=&datafile;by id;run;
data use;
       set    &datafile;
       obs=_N_;
run;
data use;
       set use;
       do i =1 to &N;
              delete_obs=0;
              if obs=i then delete_obs=1;
       output;
       end;
run;
proc sort data=use;by i;run;
data use;
       set use;
       if delete_obs=1 then delete;
run;
data use;
       set use;
       drop delete_obs;
       rename i=sampling_group;
run;

*creating empty dataset to put MLEs in;
data info;
       do sampling_group =1 to &N;
       output;
end;
run;

%DO i=1 %to &N;
*create separate datasets because proc traj doesnt allow where statement;
data use&i;
```

```
       set use;
       where sampling_group=&i;
run;
PROC TRAJ DATA=use&i OUTPLOT=OP OUTSTAT=OS OUT=OF OUTEST=OE ITDETAIL;
    ID ID; VAR &y1 &y2 &y3 &y4 &y5 &y6; INDEP &time1 &time2 &time3 &time4 &time5
&time6;
       MODEL CNORM; min &min; MAX &max; ngroups &ngroups; ORDER &order;
RUN;

*Find MLE of betas: in OE dataset;
data oe;
       set oe;
       where _type_="PARMS";
       drop _model_ _model2_ _name_;
       sampling_group=&i;
run;
data info;
       merge info oe(drop=_type_);
       by sampling_group;
run;
%end;

proc means data=info noprint;
       var _BIC1_ _BIC2_ _AIC_;
       output out=bic_means;
run;
data bic_means;
       set bic_means;
       where _STAT_="MEAN";
       rename _BIC1_=avg_BIC1_ _BIC2_=avg_BIC2_ _AIC_=avg_AIC_;
run;

proc sort data=&datafile;by id;run;
data use_;
       set &datafile;
       sampling_group=_N_;
run;
proc sort data=use_;by sampling_group;run;
proc sort data=info;by sampling_group;run;
data info;
       merge use_ info;
       by sampling_group;
       theta1=0;
       sum_exptheta=0;
       check_pi=0;
run;

%DO k=1 %to &ngroups;
data info;
       set info;
       sum_exptheta=sum_exptheta+exp(theta&k);
run;
%end;
```

```
%DO k=1 %to &ngroups;
data info;
       set info;
       mu_i_j1_k&k=sum(interc&k,&time1*linear&k,&time1*&time1*QUADRA&k);
       mu_i_j2_k&k=sum(interc&k,&time2*linear&k,&time2*&time2*QUADRA&k);
       mu_i_j3_k&k=sum(interc&k,&time3*linear&k,&time3*&time3*QUADRA&k);
       mu_i_j4_k&k=sum(interc&k,&time4*linear&k,&time4*&time4*QUADRA&k);
       mu_i_j5_k&k=sum(interc&k,&time5*linear&k,&time5*&time5*QUADRA&k);
       mu_i_j6_k&k=sum(interc&k,&time6*linear&k,&time6*&time6*QUADRA&k);
       pi_hat_k&k=exp(theta&k)/sum_exptheta;
       check_pi=check_pi+pi_hat_k&k;
run;
%end;

proc sort data=info;by id;run;
proc transpose data=info out=info_long;
       by id;
run;
data info_long;
       set info_long;
       name=upcase(_name_);
run;

data true_values;
       set info_long;
       where name="&outcome1" | name="&outcome2" | name="&outcome3" | name="&outcome4"
| name="&outcome5" | name="&outcome6";
       rename col1=outcome;
run;

%DO j=1 %to &num_timepts;
data true_values;
       set true_values;
       if index(name,"&outcome1") ge 1 then time=1;
       if index(name,"&outcome2") ge 1 then time=2;
       if index(name,"&outcome3") ge 1 then time=3;
       if index(name,"&outcome4") ge 1 then time=4;
       if index(name,"&outcome5") ge 1 then time=5;
       if index(name,"&outcome6") ge 1 then time=6;
run;
%end;

%DO k=1 %to &ngroups;
data info_muk&k;
       set info_long;
       where name contains "MU";
       if index(name,"K&k") ge 1;
       rename col1=mu&k;
run;
data info_pik&k;
       set info_long;
       where name contains "PI_HAT";
       if index(name,"K&k") ge 1;
       rename col1=pi&k;
run;
%end;
```

```
%DO j=1 %to &num_timepts;
%DO k=1 %to &ngroups;
data info_muk&k;
       set info_muk&k;
       if index(name,"J&j") ge 1 then time=&j;
run;
%end;
%end;

proc sort data=info_long nodupkeys;by id;run;
data info_long;
       set info_long;
       keep id;
run;

*creating empty dataset to put MLEs in;
proc sort data=info_long;by id;run;
data info_long;
       SET info_long;
       do time=1 to &num_timepts;
       by id;
       output;
end;
run;

%DO k=1 %to &ngroups;
proc sort data=info_long;by id time;run;
proc sort data=true_values;by id time;run;
proc sort data=info_muk&k;by id time;run;
data info_long;
       merge info_long true_values(keep=id time outcome) info_muk&k(keep=id mu&k
time);
       by id time;
run;
%end;
%DO k=1 %to &ngroups;
proc sort data=info_long;by id;run;
proc sort data=info_pik&k;by id;run;
data info_long;
       merge info_long info_pik&k(keep=id pi&k);
       by id;
       y_hat_sum=0;
run;
%end;

%DO k=1 %to &ngroups;
data info_long;
       set info_long;
       y_hat&k=(pi&k*mu&k);
       y_hat_sum=sum(y_hat_sum,y_hat&k);
run;
%end;
data info_long;
       set info_long;
       yhat=(1/&ngroups)*y_hat_sum;
       abs_dev=abs(outcome-yhat);
       MSE=(outcome-yhat)**2;
run;
proc means data=info_long noprint;
       var abs_dev MSE;
       class id;
       output out=diff_mean;
run;
```

```
data diff_mean;
       set diff_mean;
       where _Stat_="MEAN";
       if id=. then delete;
run;
proc means data=diff_mean noprint;
       var abs_dev mse;
       output out=grand_mean;
run;
data grand_mean;
       set grand_mean;
       where _Stat_="MEAN";
run;

*what BIC gives you for all available data;
PROC TRAJ DATA=&datafile OUTPLOT=OP OUTSTAT=OS OUT=OF OUTEST=OE ITDETAIL;
    ID ID; VAR &y1 &y2 &y3 &y4 &y5 &y6; INDEP &time1 &time2 &time3 &time4 &time5
&time6;
       MODEL CNORM; min &min; MAX &max; ngroups &ngroups; ORDER &order;
RUN;
data oe_full;
       set oe;
       where _type_="PARMS";
       keep _type_ _LOGLIK_--_AIC_;
       rename _LOGLIK_=LOGLIK_full_model _BIC1_=BIC1_full_model _BIC2_=BIC2_full_model
_AIC_=AIC_full_model;
run;

data info_yhat;
       merge grand_mean(keep=abs_dev mse) oe_full(drop=_type_)
bic_means(keep=avg_BIC1_
              avg_BIC2_ avg_AIC_);
       poly_order="&order";
       number_of_groups="&ngroups";
       label LOGLIK_full_model="Log Likelihood" BIC1_full_model="BIC1"
              BIC2_full_model="BIC2" AIC_full_model="AIC"
              number_of_groups="Number of Latent Classes" abs_dev="Absolute deviation
              between obs and pred" MSE="MSE between obs and pred"
              poly_order="Polynomial Degree for each Group"
              avg_BIC1_="Average BIC1 across 5 folds" avg_BIC2_="Average BIC2 across 5
              folds" avg_AIC_="Average AIC across 5 folds";
run;
proc print data=info_yhat noobs label;
       var number_of_groups poly_order ABS_DEV MSE avg_BIC1_ avg_BIC2_ avg_AIC_
BIC1_full_model BIC2_full_model AIC_full_model;
run;

%mend nfold;
```

B.2 FIVE-FOLD CV

```
%macro fivefold(datafile=, N=, ngroups=, order=, num_timepts=, y1=, y2=, y3=, y4=,
y5=, y6=,time1=,time2=,time3=,time4=,time5=,time6=,min=,max=);

%let outcome1=%upcase(&y1);
%let outcome2=%upcase(&y2);
%let outcome3=%upcase(&y3);
%let outcome4=%upcase(&y4);
%let outcome5=%upcase(&y5);
%let outcome6=%upcase(&y6);

proc sort data=&datafile;by id;run;
data use;
       set &datafile;
       obs=_N_;
run;

%let rand_samp_size=%eval(&N/5);

*First fold;
proc surveyselect data=use method=SRS rep=1
  sampsize=&rand_samp_size seed=12345 out=rand_select1;
  id _all_;
run;
data rand_select1;
       set rand_select1;
       delete=1;
       rand_samp=1;
run;
proc sort data=use;by id;run;
proc sort data=rand_select1;by id;run;
data round2;
       merge use rand_select1(keep=id delete);
       by id;
run;
data round2;
       set round2;
       where delete NE 1;
run;

*Second fold;
proc surveyselect data=round2 method=SRS rep=1
  sampsize=&rand_samp_size seed=12345 out=rand_select2;
  id _all_;
run;
data rand_select2;
       set rand_select2;
       delete=1;
       rand_samp=2;
run;
proc sort data=round2;by id;run;
proc sort data=rand_select2;by id;run;
```

```
data round3;
       merge round2 rand_select2(keep=id delete);
       by id;
run;
data round3;
       set round3;
       where delete NE 1;
run;

*Third fold;
proc surveyselect data=round3 method=SRS rep=1
  sampsize=&rand_samp_size seed=12345 out=rand_select3;
  id _all_;
run;
data rand_select3;
       set rand_select3;
       delete=1;
       rand_samp=3;
run;
proc sort data=round3;by id;run;
proc sort data=rand_select3;by id;run;
data round4;
       merge round3 rand_select3(keep=id delete);
       by id;
run;
data round4;
       set round4;
       where delete NE 1;
run;

*Fourth fold;
proc surveyselect data=round4 method=SRS rep=1
  sampsize=&rand_samp_size seed=12345 out=rand_select4;
  id _all_;
run;
data rand_select4;
       set rand_select4;
       delete=1;
       rand_samp=4;
run;
proc sort data=round4;by id;run;
proc sort data=rand_select4;by id;run;
data round5;
       merge round4 rand_select4(keep=id delete);
       by id;
run;
data round5;
       set round5;
       where delete NE 1;
run;

*Fifth fold;
data round5;
       set round5;
       rand_samp=5;
run;
proc sort data=use;by id;run;
proc sort data=rand_select1;by id;run;
proc sort data=rand_select2;by id;run;
proc sort data=rand_select3;by id;run;
proc sort data=rand_select4;by id;run;
proc sort data=round5;by id;run;
```

```
data rand_samp;
       merge use rand_select1(keep=id rand_samp) rand_select2(keep=id rand_samp)
rand_select3(keep=id rand_samp)
              rand_select4(keep=id rand_samp) round5(keep=id rand_samp);
       by id;
run;
*check;
proc freq data=rand_samp;
       tables rand_samp;
run;

*creating empty dataset to put MLEs in;
data info;
       do rand_samp=1 to 5;
       output;
end;
run;

%DO i=1 %to 5;
*create separate datasets because proc traj doesnt allow where statement;
data use_without_&i;
       set rand_samp;
       if rand_samp=&i then delete;
run;
PROC TRAJ DATA=use_without_&i OUTPLOT=OP OUTSTAT=OS OUT=OF OUTEST=OE ITDETAIL;
    ID ID; VAR &y1 &y2 &y3 &y4 &y5 &y6; INDEP &time1 &time2 &time3 &time4 &time5
&time6;
       MODEL CNORM; min &min; MAX &max; ngroups &ngroups; ORDER &order;
RUN;

*Find MLE of betas: in OE dataset;
data oe;
       set oe;
       where _type_="PARMS";
       drop _model_ _model2_ _name_;
       rand_samp=&i;
run;
data info;
       merge info oe(drop=_type_);
       by rand_samp;
run;

%end;

proc means data=info noprint;
       var _BIC1_ _BIC2_ _AIC_;
       output out=bic_means;
run;
data bic_means;
       set bic_means;
       where _STAT_="MEAN";
       rename _BIC1_=avg_BIC1_ _BIC2_=avg_BIC2_ _AIC_=avg_AIC_;
run;

proc sort data=rand_samp;by rand_samp;run;
proc sort data=info;by rand_samp;run;
data info;
       merge rand_samp info;
       by rand_samp;
       theta1=0;
       sum_exptheta=0;
       check_pi=0;
run;
```

```
%DO k=1 %to &ngroups;
data info;
       set info;
       sum_exptheta=sum_exptheta+exp(theta&k);
run;
%end;

%DO k=1 %to &ngroups;
data info;
       set info;
       mu_i_j1_k&k=sum(interc&k,&time1*linear&k,&time1*&time1*QUADRA&k);
       mu_i_j2_k&k=sum(interc&k,&time2*linear&k,&time2*&time2*QUADRA&k);
       mu_i_j3_k&k=sum(interc&k,&time3*linear&k,&time3*&time3*QUADRA&k);
       mu_i_j4_k&k=sum(interc&k,&time4*linear&k,&time4*&time4*QUADRA&k);
       mu_i_j5_k&k=sum(interc&k,&time5*linear&k,&time5*&time5*QUADRA&k);
       mu_i_j6_k&k=sum(interc&k,&time6*linear&k,&time6*&time6*QUADRA&k);
       pi_hat_k&k=exp(theta&k)/sum_exptheta;
       check_pi=check_pi+pi_hat_k&k;
run;
%end;

proc sort data=info;by id;run;
proc transpose data=info out=info_long;
       by id;
run;
data info_long;
       set info_long;
       name=upcase(_name_);
run;

data true_values;
       set info_long;
       where name="&outcome1" | name="&outcome2" | name="&outcome3" | name="&outcome4"
| name="&outcome5" | name="&outcome6";
       rename col1=outcome;
run;

%DO j=1 %to &num_timepts;
data true_values;
       set true_values;
       if index(name,"&outcome1") ge 1 then time=1;
       if index(name,"&outcome2") ge 1 then time=2;
       if index(name,"&outcome3") ge 1 then time=3;
       if index(name,"&outcome4") ge 1 then time=4;
       if index(name,"&outcome5") ge 1 then time=5;
       if index(name,"&outcome6") ge 1 then time=6;
run;
%end;
%DO k=1 %to &ngroups;
data info_muk&k;
       set info_long;
       where name contains "MU";
       if index(name,"K&k") ge 1;
       rename col1=mu&k;
run;
data info_pik&k;
       set info_long;
       where name contains "PI_HAT";
       if index(name,"K&k") ge 1;
       rename col1=pi&k;
run;
%end;
```

```
%DO j=1 %to &num_timepts;
%DO k=1 %to &ngroups;
data info_muk&k;
       set info_muk&k;
       if index(name,"J&j") ge 1 then time=&j;
run;
%end;
%end;

proc sort data=info_long nodupkeys;by id;run;
data info_long;
       set info_long;
       keep id;
run;

*creating empty dataset to put MLEs in;
proc sort data=info_long;by id;run;
data info_long;
       SET info_long;
       do time=1 to &num_timepts;
       by id;
       output;
end;
run;

%DO k=1 %to &ngroups;
proc sort data=info_long;by id time;run;
proc sort data=true_values;by id time;run;
proc sort data=info_muk&k;by id time;run;
data info_long;
       merge info_long true_values(keep=id time outcome) info_muk&k(keep=id mu&k
time);
       by id time;
run;
%end;

%DO k=1 %to &ngroups;
proc sort data=info_long;by id;run;
proc sort data=info_pik&k;by id;run;
data info_long;
       merge info_long info_pik&k(keep=id pi&k);
       by id;
       y_hat_sum=0;
run;
%end;

%DO k=1 %to &ngroups;
data info_long;
       set info_long;
       y_hat&k=(pi&k*mu&k);
       y_hat_sum=sum(y_hat_sum,y_hat&k);
run;
%end;

data info_long;
       set info_long;
       yhat=(1/&ngroups)*y_hat_sum;
       abs_dev=abs(outcome-yhat);
       MSE=(outcome-yhat)**2;
run;
```

```
proc means data=info_long noprint;
       var abs_dev MSE;
       class id;
       output out=diff_mean;
run;
data diff_mean;
       set diff_mean;
       where _Stat_="MEAN";
       if id=. then delete;
run;

proc means data=diff_mean noprint;
       var abs_dev mse;
       output out=grand_mean;
run;
data grand_mean;
       set grand_mean;
       where _Stat_="MEAN";
run;

*what BIC gives you for all available data;
PROC TRAJ DATA=&datafile OUTPLOT=OP OUTSTAT=OS OUT=OF OUTEST=OE ITDETAIL;
    ID ID; VAR &y1 &y2 &y3 &y4 &y5 &y6; INDEP &time1 &time2 &time3 &time4 &time5
&time6;
       MODEL CNORM; min &min; MAX &max; ngroups &ngroups; ORDER &order;
RUN;
data oe_full;
       set oe;
       where _type_="PARMS";
       keep _type_ _LOGLIK_--_AIC_;
       rename _LOGLIK_=LOGLIK_full_model _BIC1_=BIC1_full_model _BIC2_=BIC2_full_model
_AIC_=AIC_full_model;
run;

data info_yhat;
       merge grand_mean(keep=abs_dev mse) oe_full(drop=_type_)
bic_means(keep=avg_BIC1_ avg_BIC2_ avg_AIC_);
       poly_order="&order";
       number_of_groups="&ngroups";
       label LOGLIK_full_model="Log Likelihood" BIC1_full_model="BIC1"
              BIC2_full_model="BIC2" AIC_full_model="AIC"
              number_of_groups="Number of Latent Classes" abs_dev="Absolute deviation
              between obs and pred" MSE="MSE between obs and pred"
              poly_order="Polynomial Degree for each Group"
              avg_BIC1_="Average BIC1 across 5 folds" avg_BIC2_="Average BIC2 across 5
              folds" avg_AIC_="Average AIC across 5 folds";
run;
proc print data=info_yhat noobs label;
       var number_of_groups poly_order ABS_DEV MSE avg_BIC1_ avg_BIC2_ avg_AIC_
BIC1_full_model BIC2_full_model AIC_full_model;
run;

%mend fivefold;
```

B.3 FIVE BY TWO-FOLD CV

```
%macro fivextwofold(datafile=, seed1=, seed2=, seed3=, seed4=, seed5=, N=, ngroups=,
order=,num_timepts=,y1=,y2=,y3=,y4=,y5=,y6=,time1=,time2=,time3=,time4=,time5=,time6=,
min=,max=);

%let outcome1=%upcase(&y1);
%let outcome2=%upcase(&y2);
%let outcome3=%upcase(&y3);
%let outcome4=%upcase(&y4);
%let outcome5=%upcase(&y5);
%let outcome6=%upcase(&y6);

proc sort data=&datafile;by id;run;
data use;
       set    &datafile;
       obs=_N_;
run;

%let rand_samp_size=%eval(&N/2);

%DO REPEAT_CV=1 %to 5;

*First fold;
proc surveyselect data=use method=SRS rep=1
  sampsize=&rand_samp_size seed=&&seed&REPEAT_CV out=rand_select1;
  id _all_;
run;
data rand_select1;
       set rand_select1;
       delete=1;
       rand_samp=1;
run;
proc sort data=use;by id;run;
proc sort data=rand_select1;by id;run;
data round2;
       merge use rand_select1(keep=id delete);
       by id;
run;
data round2;
       set round2;
       where delete NE 1;
run;

*Second fold;
data round2;
       set round2;
       rand_samp=2;
run;
proc sort data=use;by id;run;
proc sort data=rand_select1;by id;run;
proc sort data=round2;by id;run;
data rand_samp;
       merge use rand_select1(keep=id rand_samp) round2(keep=id rand_samp);
       by id;
run;
*check;
proc freq data=rand_samp;
       tables rand_samp;
run;
```

```
*creating empty dataset to put MLEs in;
data info;
       do rand_samp=1 to 2;
       output;
end;
run;
%DO i=1 %to 2;
*create separate datasets because proc traj doesnt allow where statement;
data use_without_&i;
       set rand_samp;
       if rand_samp=&i then delete;
run;
PROC TRAJ DATA=use_without_&i OUTPLOT=OP OUTSTAT=OS OUT=OF OUTEST=OE ITDETAIL;
    ID ID; VAR &y1 &y2 &y3 &y4 &y5 &y6; INDEP &time1 &time2 &time3 &time4 &time5
&time6;
       MODEL CNORM; min &min; MAX &max; ngroups &ngroups; ORDER &order;
RUN;
*Find MLE of betas: in OE dataset;
data oe;
       set oe;
       where _type_="PARMS";
       drop _model_ _model2_ _name_;
       rand_samp=&i;
run;
data info;
       merge info oe(drop=_type_);
       by rand_samp;
run;

%end;
proc means data=info;
       var _BIC1_ _BIC2_ _AIC_;
       output out=bic_means;
run;
data bic_means;
       set bic_means;
       where _STAT_="MEAN";
       rename _BIC1_=avg_BIC1_ _BIC2_=avg_BIC2_ _AIC_=avg_AIC_;
run;
proc sort data=rand_samp;by rand_samp;run;
proc sort data=info;by rand_samp;run;
data info;
       merge rand_samp info;
       by rand_samp;
       theta1=0;
       sum_exptheta=0;
       check_pi=0;
run;

%DO k=1 %to &ngroups;
data info;
       set info;
       sum_exptheta=sum_exptheta+exp(theta&k);
run;
%end;
```

```
%DO k=1 %to &ngroups;
data info;
       set info;
       mu_i_j1_k&k=sum(interc&k,&time1*linear&k,&time1*&time1*QUADRA&k);
       mu_i_j2_k&k=sum(interc&k,&time2*linear&k,&time2*&time2*QUADRA&k);
       mu_i_j3_k&k=sum(interc&k,&time3*linear&k,&time3*&time3*QUADRA&k);
       mu_i_j4_k&k=sum(interc&k,&time4*linear&k,&time4*&time4*QUADRA&k);
       mu_i_j5_k&k=sum(interc&k,&time5*linear&k,&time5*&time5*QUADRA&k);
       mu_i_j6_k&k=sum(interc&k,&time6*linear&k,&time6*&time6*QUADRA&k);
       pi_hat_k&k=exp(theta&k)/sum_exptheta;
       check_pi=check_pi+pi_hat_k&k;
run;
%end;

proc sort data=info;by id;run;
proc transpose data=info out=info_long;
       by id;
run;
data info_long;
       set info_long;
       name=upcase(_name_);
run;

data true_values;
       set info_long;
       where name="&outcome1" | name="&outcome2" | name="&outcome3" | name="&outcome4"
| name="&outcome5" | name="&outcome6";
       rename col1=outcome;
run;
%DO j=1 %to &num_timepts;
data true_values;
       set true_values;
       if index(name,"&outcome1") ge 1 then time=1;
       if index(name,"&outcome2") ge 1 then time=2;
       if index(name,"&outcome3") ge 1 then time=3;
       if index(name,"&outcome4") ge 1 then time=4;
       if index(name,"&outcome5") ge 1 then time=5;
       if index(name,"&outcome6") ge 1 then time=6;
run;
%end;

%DO k=1 %to &ngroups;
data info_muk&k;
       set info_long;
       where name contains "MU";
       if index(name,"K&k") ge 1;
       rename col1=mu&k;
run;
data info_pik&k;
       set info_long;
       where name contains "PI_HAT";
       if index(name,"K&k") ge 1;
       rename col1=pi&k;
run;
%end;
%DO j=1 %to &num_timepts;
%DO k=1 %to &ngroups;
data info_muk&k;
       set info_muk&k;
       if index(name,"J&j") ge 1 then time=&j;
run;
%end;
%end;
```

```
proc sort data=info_long nodupkeys;by id;run;
data info_long;
       set info_long;
       keep id;
run;

*creating empty dataset to put MLEs in;
proc sort data=info_long;by id;run;
data info_long;
       SET info_long;
       do time=1 to &num_timepts;
       by id;
       output;
end;
run;

%DO k=1 %to &ngroups;
proc sort data=info_long;by id time;run;
proc sort data=true_values;by id time;run;
proc sort data=info_muk&k;by id time;run;
data info_long;
       merge info_long true_values(keep=id time outcome) info_muk&k(keep=id mu&k
time);
       by id time;
run;
%end;

%DO k=1 %to &ngroups;
proc sort data=info_long;by id;run;
proc sort data=info_pik&k;by id;run;
data info_long;
       merge info_long info_pik&k(keep=id pi&k);
       by id;
       y_hat_sum=0;
run;
%end;

%DO k=1 %to &ngroups;
data info_long;
       set info_long;
       y_hat&k=(pi&k*mu&k);
       y_hat_sum=sum(y_hat_sum,y_hat&k);
run;
%end;

data info_long;
       set info_long;
       yhat=(1/&ngroups)*y_hat_sum;
       abs_dev=abs(outcome-yhat);
       MSE=(outcome-yhat)**2;
run;
proc means data=info_long noprint;
       var abs_dev MSE;
       class id;
       output out=diff_mean;
run;
data diff_mean;
       set diff_mean;
       where _Stat_="MEAN";
       if id=. then delete;
run;
```

```
proc means data=diff_mean noprint;
       var abs_dev mse;
       output out=grand_mean;
run;
data grand_mean;
       set grand_mean;
       where _Stat_="MEAN";
run;

data info_yhat&REPEAT_CV;
       merge grand_mean(keep=abs_dev mse) bic_means(keep=avg_BIC1_ avg_BIC2_
avg_AIC_);
       poly_order="&order";
       number_of_groups="&ngroups";
       round=&REPEAT_CV;
       label number_of_groups="Number of Latent Classes" abs_dev="Absolute deviation
              between obs and pred" MSE="MSE between obs and pred"
              poly_order="Polynomial Degree for each Group"
              avg_BIC1_="Average BIC1 across 5 folds" avg_BIC2_="Average BIC2 across 5
              folds" avg_AIC_="Average AIC across 5 folds";
run;
%end;

*creating empty dataset to put info in;
data info_yhat;
       do round=1 to 5;
       output;
end;
run;

%do REPEAT_CV=1 %to 5;
data info_yhat;
       merge info_yhat info_yhat&REPEAT_CV;
       by round;
run;
%end;

PROC MEANS DATA=INFO_YHAT;
       var ABS_DEV MSE avg_BIC1_ avg_BIC2_ avg_AIC_;
run;

%MEND FIVEXTWOFOLD;
```

BIBLIOGRAPHY

1. Loughran, T. and D.S. Nagin, *Finite Sample Effects in Group-Based Trajectory Models.* Sociological Methods & Research, 2006. **35**(2): p. 250-278.
2. Blokland, A.A.J., D. Nagin, and P. Nieuwbeerta, *Life Span Offending Trajectories of a Dutch Conviction Cohort.* Criminology, 2005. **43**(4): p. 919-954.
3. D'Unger, A., K. Land, and P. McCall, *Sex Differences in Age Patterns of Delinquent/Criminal Careers: Results from Poisson Latent Class Analyses of the Philadelphia Cohort Study.* Journal of Quantitative Criminology, 2002. **18**(4): p. 349-375.
4. Piquero, A.R., et al., *Assessing the Impact of Exposure Time and Incapacitation on Longitudinal Trajectories of Criminal Offending.* Journal of Adolescent Research, 2001. **16**(1): p. 54-74.
5. Yessine, A.K. and J. Bonta, *The Offending Trajectories of Youthful Aboriginal Offenders&lt;sup&gt;1&lt;/sup&gt.* Canadian Journal of Criminology and Criminal Justice/La Revue canadienne de criminologie et de justice pénale, 2009. **51**(4): p. 435-472.
6. Nagin, D.S., *Analyzing developmental trajectories: a semiparametric, group-based approach.* Psychological methods, 1999. **4**(2): p. 139.
7. Refaeilzadeh, P., L. Tang, and H. Liu. *On comparison of feature selection algorithms*. in *Proceedings of AAAI Workshop on Evaluation Methods for Machine Learning II*. 2007.
8. Nielsen, J., et al., *Group-based Criminal Trajectory Analysis using Cross-Validation Criteria*. 2011.
9. Nagin, D.S. and R.E. Tremblay, *What Has Been Learned from Group-Based Trajectory Modeling? Examples from Physical Aggression and Other Problem Behaviors.* Annals of the American Academy of Political and Social Science, 2005. **602**: p. 82-117.
10. Bryk, A.S. and S.W. Raudenbush, *Application of hierarchical linear models to assessing change.* Psychological Bulletin, 1987. **101**(1): p. 147-158.
11. Bryk, A.S. and S.W. Raudenbush, *Hierarchical linear models: applications and data analysis methods*. Vol. 1. 1992, Newbury Park: Sage Publications.
12. Goldstein, H., *Multilevel statistical models*. Vol. 922. 2011: John Wiley & Sons.
13. McArdle, J.J. and D. Epstein, *Latent Growth Curves within Developmental Structural Equation Models.* Child Development, 1987. **58**(1): p. 110-133.
14. Meredith, W. and J. Tisak, *Latent curve analysis.* Psychometrika, 1990. **55**(1): p. 107-122.
15. Muthén, B.O., *Latent variable modeling in heterogeneous populations.* Psychometrika, 1989. **54**(4): p. 557-585.
16. Willett, J.B. and A.G. Sayer, *Using covariance structure analysis to detect correlates and predictors of individual change over time.* Psychological Bulletin, 1994. **116**(2): p. 363.

17. Jones, B.L., D.S. Nagin, and K. Roeder, *A SAS Procedure Based on Mixture Models for Estimating Developmental Trajectories.* Sociological Methods & Research, 2001. **29**(3): p. 374-393.

18. Nagin, D.S. and K.C. Land, *Age, Criminal Carrers, and Population Heterogeneity: Specification and Estimation of a Nonparametric, Mixed Poisson Model.* Criminology, 1993. **31**(3): p. 327-362.

19. Roeder, K., K.G. Lynch, and D.S. Nagin, *Modeling Uncertainty in Latent Class Membership: A Case Study in Criminology.* Journal of the American Statistical Association, 1999. **94**(447): p. 766-776.

20. Nagin, D.S. and K.C. Land, *Age, Criminal Careers, and Population Heterogeneity: Specification and Estimation of a Nonparametric, Mixed Poisson Model.* Criminology, 1993. **31**(3): p. 327-362.

21. Land, K.C., P.L. McCall, and D.S. Nagin, *A Comparison of Poisson, Negative Binomial, and Semiparametric Mixed Poisson Regression Models: With Empirical Applications to Criminal Careers Data.* Sociological Methods & Research, 1996. **24**(4): p. 387-442.

22. Nagin, D. and R.E. Tremblay, *Trajectories of Boys' Physical Aggression, Opposition, and Hyperactivity on the Path to Physically Violent and Nonviolent Juvenile Delinquency.* Child Development, 1999. **70**(5): p. 1181-1196.

23. Nagin, D.S., *Group-Based Modeling of Development*. 2005, Cambridge, MA, USA: Harvard University Press.

24. Schwarz, G., *Estimating the Dimension of a Model.* The Annals of Statistics, 1978. **6**(2): p. 461-464.

25. Larson, S.C., *The shrinkage of the coefficient of multiple correlation.* Journal of Educational Psychology, 1931. **22**(1): p. 45.

26. Mosteller, F. and J.W. Tukey, *Data analysis, including statistics*. 1968.

27. Ghosh, J.K. and P.K. Sen, *On the asymptotic performance of the log likelihood ratio statistic for the mixture model and related results*. 1984, Citeseer.

28. D'Unger, A., et al., *How Many Latent Classes of Delinquent/Criminal Careers? Results from Mixed Poisson Regression Analyses.* American Journal of Sociology, 1998. **103**(6): p. 1593-1630.

29. Ota, R., et al., *Appropriate likelihood ratio tests and marginal distributions for evolutionary tree models with constraints on parameters.* Molecular Biology and Evolution, 2000. **17**(5): p. 798-803.

30. Wasserman, L., *Bayesian model selection and model averaging.* Journal of mathematical psychology, 2000. **44**(1): p. 92-107.

31. Eggleston, E., J. Laub, and R. Sampson, *Methodological Sensitivities to Latent Class Analysis of Long-Term Criminal Trajectories.* Journal of Quantitative Criminology, 2004. **20**(1): p. 1-26.

32. Dietterich, T.G., *Approximate statistical tests for comparing supervised classification learning algorithms.* Neural computation, 1998. **10**(7): p. 1895-1923.

33. Franzen, P.L., et al., *Poor sleep quality predicts onset of either major depression or subsyndromal depression with irritability during interferon-alpha treatment* Journal of Psychiatric Research, 2009. **e-pub ahead of print; doi:10.1016/j.psychres.2009.02.011.**

34. First, M.B., et al., *Structured Clinical Interview for DSM-IV Axis I Disorders, Patient Edition, January 1995 FINAL*. 1995, SCID-I/P Version 2.0). New York, NY: Biometrics Research Department, New York State Psychiatric Institute.

35. Miller, M.D., et al., *Rating chronic medical illness burden in geropsychiatric practice and research: application of the Cumulative Illness Rating Scale.* Psychiatry research, 1992. **41**(3): p. 237-248.

36. Beck, A.T., et al., *Comparison of Beck Depression Inventories-IA and-II in psychiatric outpatients.* Journal of personality assessment, 1996. **67**(3): p. 588-597.

37. Buysse, D.J., et al., *The Pittsburgh Sleep Quality Index: a new instrument for psychiatric practice and research.* Psychiatry research, 1989. **28**(2): p. 193-213.

38. Backhaus, J., et al., *Test–retest reliability and validity of the Pittsburgh Sleep Quality Index in primary insomnia.* Journal of psychosomatic research, 2002. **53**(3): p. 737-740.

39. Hall, M., et al., *Symptoms of stress and depression as correlates of sleep in primary insomnia.* Psychosomatic Medicine, 2000. **62**(2): p. 227-230.