

---

[All ETDs from UAB](#)

[UAB Theses & Dissertations](#)

---

2019

## Applying Markov Processes To Model Disability In Relapsing Multiple Sclerosis Patients

Anastasia Hartzes

*University of Alabama at Birmingham*

Follow this and additional works at: <https://digitalcommons.library.uab.edu/etd-collection>

---

### Recommended Citation

Hartzes, Anastasia, "Applying Markov Processes To Model Disability In Relapsing Multiple Sclerosis Patients" (2019). *All ETDs from UAB*. 1886.

<https://digitalcommons.library.uab.edu/etd-collection/1886>

This content has been accepted for inclusion by an authorized administrator of the UAB Digital Commons, and is provided as a free open access item. All inquiries regarding this item or the UAB Digital Commons should be directed to the [UAB Libraries Office of Scholarly Communication](#).

APPLYING MARKOV PROCESSES TO MODEL DISABILITY  
IN  
RELAPSING MULTIPLE SCLEROSIS PATIENTS

by

ANASTASIA MARIA HARTZES

CHARITY J. MORGAN, COMMITTEE CHAIR  
STACEY COFIELD  
GARY CUTTER  
ELLEN EATON  
BYRON JAEGER  
JOHN RINKER

A DISSERTATION

Submitted to the faculty of The University of Alabama at Birmingham,  
in partial fulfillment of the requirements for the degree of  
Doctor of Philosophy.

BIRMINGHAM, ALABAMA

2019

© Copyright by  
Anastasia Maria Hartzes  
2019

# APPLYING MARKOV PROCESSES TO MODEL DISABILITY IN RELAPSING MULTIPLE SCLEROSIS PATIENTS

ANASTASIA HARTZES

DOCTOR OF PHILOSOPHY IN BIOSTATISTICS

## ABSTRACT

Modeling disability in multiple sclerosis (MS) is challenging due to its complexity and non-linearity, with utilized methodology having many limitations. Lack of a suitable biomarker has led to relying on statistical models to understand disease progression. Markov methodology has been limitedly applied to the clinically-assessed Expanded Disability Status Score (EDSS), but not to patient-reported Patient Determined Disease Steps (PDDS); both measure disease progression and disability, and neither have been analyzed using the Test of Lumpability (TOL). It is common practice to aggregate these scores for computational or inferential convenience; in the case of Markov Chains (MC), combining states is referred to as lumping. The resulting chain must be evaluated using the TOL to ensure retention of the Markov property. Extending the TOL, we developed a goodness of fit (GOF) test with Pearson and likelihood ratio formulations to compare lumping schemes that pass the TOL; both were shown to follow a Chi-squared distribution. Performance was evaluated using simulated and patient data.

Lumping schemes were identified for each disability scale. Using semi-annual surveys from the North American Research Committee on Multiple Sclerosis (NARCOMS), PDDS was predicted using Markov models with and without covariates

for multiple lumping schemes; using semi-annual follow-up data from the CombiRx trial, EDSS scores were similarly modeled. Disability scores were lumped with scientifically supported schemes. Schemes were assessed for parsimony, clinical usefulness, and adherence to Markov property (lumpability); covariates were selected with scientific justification.

Novel application to NARCOMS PDDS data will benefit from a larger sample size and wider range of disease statuses than are observed in clinical trials, enhancing generalizability; novel evaluation of lumpability to EDSS outcomes will extend current work. Implementation of Markov methodology has the potential to provide fresh insight into MS disease progression.

Keywords: Multiple sclerosis, Markov chains, lumpability, aggregating states, grouping states, disability prediction

## DEDICATION

This work is dedicated to my husband, Micah, who supported me with every step of this journey; and to my parents, Michael and Irene, grandparents, Anna and Tito Hartzes and Evangelia Vergos, and sister, Evangelia, who have encouraged from the start of my education, all those years ago.

## ACKNOWLEDGMENTS

This journey would not have been possible without the graciousness of a great many people. My interest in medical research began while working for Dr. George Koulianos; this led to my first real experience with research was fresh out of college as research assistant in a basic sciences lab at the Memphis, Tennessee Veteran's Affairs Hospital, where Drs. Michael Levin and Lidia Gardner were studying Multiple Sclerosis. While I loved lab work, I was left with lingering questions about what could be done with all the data they accrued during their work, and that led me on the winding road to this field. Those early experiences were invaluable to me, and I was honored to continue to study the field, and am grateful to the patients whose courage continues to drive research.

I would like to express my sincerest gratitude to my advisor, Dr. Charity Morgan, for her invaluable guidance, knowledge and patience since taking me on as a dissertation mentee; it was a privilege to have learned from and worked with her. With her enthusiasm for the field, Dr. Stacey Cofield brought me to that point because she encouraged my interest in biostatistics while I was earning my Master of Public Health in Epidemiology, and became my first advisor upon joining the department.

I am grateful to the members of my dissertation committee who have shared their expertise and time away from their busy schedules to assist me on this endeavor; and to Dr. Charles Katholi, who taught me a great deal of my theoretical courses, and who first introduced me this branch of statistics and brought instrumental insight to the early stages of this work. I am also indebted to our other professors whose dedication, encouragement and thoughtful instruction gave me the tools to make it this point.

A fruitful journey is one enriched with friends and family. I am grateful to Kelsey Jordan, Allison Fialkowski, Vincent Laufer and Tarrant McPherson, whose friendship and good humor were a bastion, especially early on, and Rouba Chahine, Satpalsinh Chandel and Stephanie Tison, whose positivity kept me energized while concluding this work; and Andrea Pappas Jernigan and Sarah Kalaris, whose support have been steadfast from the beginning. My cousin, Elektra Apostola, was a wonderful friend and sounding board, all while being halfway across the world and working on her own doctorate.

Finally, I wish to express gratitude to my family: aunts, uncles, cousins, and in-laws, especially Maria and Nick Stratas Katie and J Sutton, Nick Vergos, Christina Eady, Angela Sessions and Marlyse Gakis. Their support and understanding are indescribable, and their encouragement never wavered.



## TABLE OF CONTENTS

<b><i>ABSTRACT</i></b> .....	<b><i>iii</i></b>
<b><i>DEDICATION</i></b> .....	<b><i>v</i></b>
<b><i>ACKNOWLEDGMENTS</i></b> .....	<b><i>vi</i></b>
<b><i>LIST OF TABLES</i></b> .....	<b><i>xiv</i></b>
<b><i>LIST OF FIGURES</i></b> .....	<b><i>xvii</i></b>
<b><i>I. Background</i></b> .....	<b><i>1</i></b>
A. Disease background and public health importance .....	1
B. Data .....	2
C. Methodology .....	4
1. Current status of methodology in disability analysis .....	4
1.1. Relapse .....	4
1.2. EDSS to measure disability .....	5
1.3. PDDS to measure disability .....	5
2. State of Markov methodology in MS research .....	7
D. Explanation in support of new methodology .....	8
1. The EDSS as a primary endpoint .....	8

2.	The Nature of EDSS scores related to Markov models .....	11
3.	PDDS as a primary endpoint and Markov applicability .....	12
4.	The Nature of PDDS scores related to Markov models .....	13
E.	Gaps and Contributions .....	14
<b>II.</b>	<b><i>Theoretical background</i></b> .....	<b>17</b>
A.	Introduction and summary, notation .....	17
1.	Summary of stochastic processes and introduction of notation .....	17
2.	Summary of and introduction to Markov processes .....	17
B.	Description of Markov processes and chains .....	19
1.	Properties of Markov chains .....	19
2.	Higher order models .....	22
3.	Treatment of time .....	23
4.	Classification of states .....	25
C.	Lumpability: collapsing states to reduce the size of the transition matrix .....	28
1.	Description of lumpability .....	28
2.	Evaluating lumpability .....	30
D.	Markov (transition) models .....	32
1.	Using covariates to predict state transitions .....	32
2.	Higher order models .....	35

E.	Context of Markov models and multiple sclerosis.....	36
1.	Distribution of EDSS.....	36
2.	Distribution of PDDS.....	37
3.	Relevance of lumpability to EDSS and PDDS.....	37
F.	Model selection and Goodness of Fit: previous methods, proposed considerations .....	38
1.	Model selection to estimate parameters (transition probabilities).....	38
2.	Goodness of fit.....	39
<b>III.</b>	<b><i>Foundational theoretical work and development of a Goodness of Fit Test</i></b> <b>41</b>	
A.	Introduction.....	41
B.	Methods.....	41
1.	Description of simulated data and states .....	41
2.	Evaluation of lumpability.....	42
3.	Proposed tests: exploration of the goodness of fit test statistics and degrees of freedom .....	43
3.1	Likelihood estimation.....	44
3.2	Pearson test formulation .....	45
3.3	W-score.....	46

3.4	Simulation methodology .....	47
4.	Linear Algebra.....	48
4.1	Implementation of a lumping scheme and performing the test of lumpability .....	48
4.2	Goodness of fit test statistics .....	55
4.2.1	Likelihood ratio Goodness of Fit Test .....	55
4.2.2	Pearson Goodness of Fit Test.....	58
4.2.3	The $W$ -score .....	60
C.	Results.....	64
1.	Summary statistics .....	64
2.	Distribution of test statistics and degrees of freedom .....	64
D.	Discussion.....	70
1.	Goodness of Fit.....	70
2.	$W$ -score .....	70
3.	Performance of proposed goodness of fit tests.....	73
3.1	Type I error .....	73
3.2	Power .....	73
3.3	Comparing the Chains .....	75

E.	Summary, Discussion and Conclusion .....	79
<b>IV.</b>	<b><i>Applications to NARCOMS Registry Data: PDDS and Lumpability .....</i></b>	<b>81</b>
A.	Introduction .....	81
B.	Methods .....	81
1.	Study Design .....	81
2.	Preliminary Analysis.....	83
3.	Primary Analysis .....	83
4.	Secondary Analysis .....	85
C.	Results .....	86
1.	Preliminary Analyses.....	86
2.	Primary Analyses .....	91
3.	Secondary Analyses.....	99
D.	Discussion and Conclusion.....	104
<b>V.</b>	<b><i>Applications to the CombiRx Trial Data: EDSS and Lumpability.....</i></b>	<b>110</b>
A.	Introduction .....	110
B.	Methods .....	110
1.	Study Design .....	110
2.	Primary Analyses .....	112
3.	Secondary Analyses.....	114

C. Results.....	117
1. Preliminary Analyses.....	117
2. Primary Analyses .....	119
123	
3. Secondary Analyses.....	124
D. Discussion and Conclusion .....	126
<b>VI. Summary, conclusion, next steps.....</b>	<b>130</b>
A. Implementation of the Test of Lumpability and Novel test development and exploration.....	131
B. Comparison of NARCOMS and CombiRx results .....	132
C. Conclusion .....	137
<b>VII. Index of notation.....</b>	<b>139</b>
<b>VIII. References.....</b>	<b>140</b>
<b>APPENDIX.....</b>	<b>149</b>

## LIST OF TABLES

<b>Table 1:</b> Correspondence between PDDS and EDSS disability scales .....	13
<b>Table 2:</b> Methods of estimating transition probabilities in logistic models .....	34
<b>Table 3:</b> Transition probabilities used for simulation .....	42
<b>Table 4 :</b> Summary of general simulation steps.....	48
<b>Table 5:</b> Probabilities used for simulation.....	71
<b>Table 6:</b> Summary of proposed test statistics for Chain 1 lumpable matrices, lumpable matrices only .....	72
<b>Table 7:</b> Summary of cohort demographic characteristics at enrollment (N=2,047) .....	89
<b>Table 8:</b> Summary of cohort clinical characteristics at enrollment (N=2,047) .....	90
<b>Table 9 :</b> Unadjusted, overall transition matrices (N=2,047) .....	92
<b>Table 10:</b> Scheme 1: Unadjusted lumped transition matrix, overall (N=2,047; 9 states to 8 lumps).....	93
<b>Table 11:</b> Scheme 2: Unadjusted lumped transition matrix, overall (N=2,047; 9 states to 5 lumps).....	93
<b>Table 12:</b> Unadjusted transition matrices, Mild Impairment (N=1,545 for 16,995 transitions) .....	95

<b>Table 13:</b> Scheme 1: Unadjusted, lumped transition matrix, Mild Impairment (N=1,545; 9 states to 8 lumps) .....	95
<b>Table 14:</b> Scheme 2: Unadjusted, lumped transition matrix, Mild Impairment (N=1,545; 9 states to 5 lumps) .....	96
<b>Table 15:</b> Unadjusted transition matrices, High Impairment (N=502 for 5,522 transitions) .....	96
<b>Table 16 :</b> Scheme 1: Unadjusted, lumped transition matrix, High Impairment (N=502; 9 states to 8 lumps) .....	97
<b>Table 17:</b> Scheme 2 : Unadjusted, lumped transition matrix, High Impairment (N=502; 9 states to 5 lumps) .....	99
<b>Table 18:</b> Summary of results for complete case and stratified by enrollment disability, unadjusted probabilities .....	99
<b>Table 19:</b> Summary of proportional odds models, overall sample† .....	100
<b>Table 20:</b> Summary of proportional odds models, Mildly Impaired .....	102
<b>Table 21:</b> Summary of proportional odds models, Highly Impaired .....	103
<b>Table 22:</b> Summary of cohort demographics at enrollment or baseline (N=725) .	117
<b>Table 23:</b> Summary of cohort clinical characteristics baseline (N=725) .....	118
<b>Table 24:</b> Summary of results for unadjusted matrices, N=725 ( $\alpha=0.05$ ) .....	119
<b>Table 25:</b> Unadjusted, overall frequency transition matrix (N=725) .....	119
<b>Table 26:</b> Unadjusted, overall probability transition matrix (N=725) .....	120
<b>Table 27:</b> Scheme 1: Simple combination, unadjusted transitions (14 states to 7 lumps) .....	121



<b>Table 28:</b> Scheme 2: Baseline groupings unadjusted transitions (14 states to 4 lumps).....	121
<b>Table 29:</b> Scheme 3: PDDS Matching, unadjusted transitions (14 states to 11 lumps).....	122
<b>Table 30 :</b> Summary of proportional odds models with random effects.....	125

## LIST OF FIGURES

<b>Figure 1</b> : Basic 3-state, one-step transition matrix for a given time, $t$ .....	21
<b>Figure 2</b> : Second order, 3-state, transition matrix for a given time, $t$ .....	23
<b>Figure 3</b> : Diagram of transitions between states A, B, C, D and E in a Markov Chain.....	27
<b>Figure 4</b> : 5-State transition matrix associated with the Markov chain depicted in Figure 3 .....	28
<b>Figure 5</b> : Lumping matrices.....	29
<b>Figure 6</b> : Original example matrix expressed with counts, $3 \times 3$ dimension.....	31
<b>Figure 7</b> : Lumped matrix expressed with counts, $2 \times 2$ dimension.....	32
<b>Figure 8</b> : Lumping scheme implementation .....	43
<b>Figure 9</b> : Histograms of the Likelihood Ratio and Pearson test statistics; $N=50$ for 927 iterations .....	66
<b>Figure 10</b> : Histograms of the Likelihood Ratio and Pearson test statistics; $N=200$ for 970 iterations .....	67
<b>Figure 11</b> : Histograms of the Likelihood Ratio and Pearson test statistics; $N=1000$ for 942 iterations .....	68
<b>Figure 12</b> : Histograms of the Likelihood Ratio and Pearson test statistics; $N=5000$ for 949 iterations .....	69
<b>Figure 13</b> : Type 1 error plot, Chain 1 .....	73

<b>Figure 14:</b> Power plots, chains 2-4 .....	74
<b>Figure 15:</b> Stage 1, rejecting the test of lumpability* .....	75
<b>Figure 16:</b> Stage 2, Choosing unlumped after determining lumpability .....	76
<b>Figure 17:</b> Two-stage process, choosing lumped after both tests.....	78
<b>Figure 18:</b> Lumping schemes for the PDDS.....	86
<b>Figure 19:</b> State transition diagram for the overall sample, unlumped chain .....	94
<b>Figure 20:</b> State transition diagram for lumped chain according to Scheme 1, High Impairment.....	98
<b>Figure 21:</b> Proposed lumping schemes for all EDSS scores.....	115
<b>Figure 22 :</b> Lumping schemes as applied to the EDSS scores from study period ...	116
<b>Figure 23 :</b> State transition diagram for lumped chain according to Scheme 3 (PDDS matching) .....	123

## ***I. Background***

### **A. Disease background and public health importance**

Multiple sclerosis (MS) is a demyelinating neurodegenerative disease facing men and women globally; it has no cure or known cause. The disease most commonly presents in adulthood, with diagnoses occurring between the ages of 20 and 50; pediatric cases are rare. Three primary types of the disease exist: relapsing remitting (RRMS), primary progressive (PPMS) and secondary progressive (SPMS). RRMS is the most commonly diagnosed. Until recently, MS with relapse (RRMS and SPMS) were the only types of MS with an approved disease-modifying treatments; in March 2017, the Federal Drug Administration approved a therapy for PPMS (F. D. Lublin et al., 2014; Montalban et al., 2017; Mulero, Midaglia, & Montalban, 2018). A chronic illness, it is characterized by periods of relapse and remission, where vision and mobility are affected. Over the course of disease progression, patients experience depression, impaired motor function and decreased mobility, particularly following a relapse; daily activities are affected over time. While other measures exist, the disease progression is typically monitored through MRI and Expanded Disability Status Scale (EDSS) and by Patient Determined Disease Steps (PDDS) (D. Goodin, 2014; D. S. Goodin et al., 2016). The MRI functions to detect changes in the presence of plaques in the white matter while the EDSS is a measure of disability and disease burden; both are clinically-based (i.e., not self-reported).

The PDDS is a validated measure that was developed as a self-report surrogate of the EDSS (Learmonth, Motl, Sandroff, Pula, & Cadavid, 2013; Marrie & Goldman, 2007). On average, MS patients receiving first-line therapies experience 1 to 2 relapses per calendar year (Roskell, Zimovetz, Rycroft, Eckert, & Tyas, 2012); however, further elucidation is necessary to understand the relationship between relapse and mobility level, and the transitions between relapse and remission (D. S. Goodin et al., 2016).

To date, modeling MS disease course has been fraught with challenges, and any varied success contains limitations (Bergamaschi & Montomoli, 2016; M. Hutchinson, 2016; Taylor, 2016). A complex disease, there is no known biomarker to indicate disease presence or disability progression; therefore, with EDSS and PDDS being the best available tools, we must devote efforts to clarifying their use in describing MS progression (F. D. Lublin et al., 2014; Taylor, 2016). To that end, this work will endeavor to fill in the gaps of existing MS literature as it pertains to EDSS and PDDS analyses via Markovian methodology. The ultimate goal is to have a clearer understanding of flow of disability over disease course and predict that disability over time.

## B. Data

Methodology will be applied to three datasets; two are existing datasets, each with two different collection mechanisms and missing data structures, and the third will be created via simulation. The first dataset is from the CombiRx trial, which has

very little missing data; all patients randomized were analyzed on the primary endpoint. CombiRx was a 3-arm, double-blind, randomized North American clinical trial investigating the combination of the established disease-modifying treatments (DMT) interferon  $\beta$  1-a (IFN) and glatiramer acetate (GA) in 1,008 patients with RRMS; patients were followed for 3 to 7 years, with enrollment beginning in 2005. Its baseline information and final study results have been published (Lindsey, et al., 2011; Lublin, et al., 2013), with the extension study results published recently (F. D. Lublin et al., 2017). The second dataset is from the North American Research Committee on Multiple Sclerosis (NARCOMS); adult ( $\geq 18$  years) males and females with any form of MS voluntarily participate in this longitudinal registry. Over 38,000 patients are enrolled, contributing to more than 20 years of data (NARCOMS, 2017). Questionnaires encompass demographic and disease-course information (Cofield, Thomas, Tyry, Fox, & Salter, 2017). Data are collected twice per year, after enrollment; because data are self-report, there is a larger number of missing observations than in clinical trials.

Foundational theoretical work was performed based on data designed to evaluate our proposed methodology. Data were simulated for methodology exploration and development and are described in Chapter III.

## C. Methodology

### 1. Current status of methodology in disability analysis

#### 1.1. Relapse

A relapse is defined as experiencing a new or worsening symptom for at least 24 hours; its occurrence must be at least 30 days from the previous relapse to be considered a separate event (Lindsey et al., 2012). Relapse can be viewed either in terms of its short-term or its long-term effects on disability, as it is believed that accumulation of relapses are associated with decline in overall mobility and neurologic function (D. S. Goodin et al., 2016). A (multivariate) Markov model was first used to analyze MS data in 1985, in an effort to model the natural history of the disease, with relapse as a primary covariate for the outcome consisting of five defined states (relapse, relapse with sequelae, progression, death, lost to follow-up) (C. Wolfson & Confavreux, 1985). To date, there have been a variety of methods employed to model relapse in MS patients including Poisson regression, negative binomial regression, logistic regression, Kaplan Meier and Cox proportional hazards models (Fred D. Lublin et al., 2013; Mieno, Yamaguchi, & Ohashi, 2011; Y. C. Wang, Meyerson, Tang, & Qian, 2009). Because of the relationship between relapse and disease progression, relapse reduction was the sole endpoint in most MS trials until the mid-1990s (Meyer-Moock, Feng, Maeurer, Dippel, & Kohlmann, 2014).

## 1.2. EDSS to measure disability

Around the time the natural history Markov model was published , the EDSS was introduced and began being utilized as a means of quantifying disease progression (Kurtzke, 1955, 1983). As mentioned earlier, relapse had been a primary endpoint in trials; however, in 1996, studies began implementing the EDSS to evaluate progression as a primary endpoint (Meyer-Moock et al., 2014).

In a short term sense (six-month increments), EDSS scores have been used to model and predict short term-disability using a partial proportional odds model (Gauthier et al., 2007). In clinical trials (like CombiRx) and in clinical settings, disability is measured largely using EDSS. Scores range from 0 to 10 in half-point increments; higher scores indicate greater disability (D. S. Goodin et al., 2016; Kurtzke, 1983). Use of this score varies between studies to investigate long-term disability, such as: creating a change score; using the threshold of maintaining a score of 6 or 7; considering confirmed progression as having an increase of 1-2 points for a period of time (D. S. Goodin et al., 2016). Further discussion of these methods and their disadvantages are explored in Chapter I, Section D (The EDSS as a Primary Endpoint).

## 1.3 PDDS to measure disability

The Patient Determined Disease Scale (PDDS) is a self-reported measure of disability in MS. Scores range from 0-8 in one-unit increments. Higher scores



indicate greater disability; scores of 0 represent “normal” mobility with no symptomatic impact on daily activity; scores of 8 indicate patient is bedridden (Hohol, Orav, & Weiner, 1999; Marrie & Goldman, 2007; Rizzo, Hadjimichael, Preiningerova, & Vollmer, 2004).

The PDDS was based on the Disease Steps (DS) created by Hohol and colleagues (Rizzo et al., 2004). The DS was developed to provide clinicians a simpler means of evaluating MS disease progression compared to the EDSS, and to be useful in the short term with the hope of demonstrating its long-term functionality, according to the authors. At the time of its introduction, the EDSS was already the conventional choice for this purpose in clinical trials (Hohol, Orav, & Weiner, 1995; Hohol et al., 1999). However, due to the complexity of applying the EDSS, the intention was for non-MS specialists to have a means of evaluating disease progression. Hohol and colleagues (1995) demonstrated strong correlation between DS and EDSS scores in the short term (22-month period); they later demonstrated this for the long term (1, 2 and 3 years) (1999). The PDDS has been validated and shown to be strongly correlated with the EDSS, although this relationship is not one-to-one (Hohol et al., 1995; Learmonth et al., 2013; Marrie & Goldman, 2007). It was created as a patient-reported outcome (PRO) by researchers affiliated with NARCOMS. Upon its creation, it was reported to have a correlation of 0.958 with the EDSS (Rizzo et al., 2004).

As an endpoint, the PDDS has been modeled most commonly using logistic regression, by dichotomously defining disability progression based on an increase in score (Cofield, Fox, Tyry, Salter, & Campagnolo, 2016; Liu et al., 2016). The mean

change of PDDS has been modeled using linear regression (Cofield et al., 2016).

Observed and categorized PDDS scores have also been modeled using ordinal and nominal logistic regression (Fitzgerald et al., 2018). The NARCOMS registry collects this outcome in every survey in order to have a measure of patient disability.

## 2. State of Markov methodology in MS research

Markov processes historically have been used in medical decision making and modeling lifetime health and life expectancy (Beck & Pauker, 1983; Regnier & Schechter, 2013; Sonnenberg & Beck, 1993). While there is a history of applying Markov processes to a variety of MS data, it is limited. The earliest use recorded in the literature was the work of C. Wolfson and Confavreux (1985) (mentioned earlier in this section), to model the natural history of MS; here, authors modeled 5 disease states based on relapse (C. Wolfson & Confavreux, 1985, 1987). Markov models have been used to model relapses, alone, but primarily in the context of cost-analysis and not widely used for clinical trial data (Mieno et al., 2011; Palace et al., 2014). They have also been used to model disease progression for decision analysis regarding first-line treatment (Bargiela et al., 2017). Markov models also have been applied to lesion count data (Altman & Petkau, 2005) and EDSS scores (Palace et al., 2014). Markov models have been predominantly used for determining the economics surrounding a particular treatment, although some authors have argued it is inappropriate to utilize the EDSS for this purpose (Fisk, et al., 2005).

Beyond these applications, Markov models have not been consistently employed to investigate various aspects of MS, of which modeling EDSS is a particular example. Specifically, it is of particular importance in MS to have a clear understanding of disability over time, which is reflected in movement between EDSS states. Finally, there is no identified evidence in the literature indicating Markov methodology has been applied to PDDS, nor has it been applied to NARCOMS data, in general.

#### D. Explanation in support of new methodology

##### 1. The EDSS as a primary endpoint

**Use, meaning, strengths, weakness** EDSS is the most commonly utilized method of evaluation for new therapies and disease progression, and is widely accepted as a strong endpoint in trials. It is used as a primary endpoint and, to date, it is the most common secondary endpoint in relapsing-MS (RMS) clinical trials (Meyer-Moock et al., 2014). Clinicians evaluate 8 functional systems of the central nervous system: vision, brainstem, pyramidal, cerebellar, sensory, bowel/bladder, cerebral, ambulation which results in a composite score reflecting disability (Baldassari, Salter, Longbrake, Cross, & Naismith, 2017; Kurtzke, 1983). Scores of 0 represent normal neurologic function, and scores of 10 indicate death to MS. Lower scores on the scale (EDSS  $\leq 5.5$ ) represent neurological impairments and higher scores (EDSS  $> 6$ ) represent disability; the intermediate scores (EDSS 4-6) are driven by mobility (Meyer-Moock et al., 2014).

Using EDSS changes as an indicator, a clinical increase in disability (disease progression) is determined if (1) a baseline score is 5.5 or less and changes by 1.0 point or (2) a baseline score is 6.0 or higher and changes by 0.5 point. Further, this score change must be maintained for 12 or 24 weeks (this range varies in the literature). Patient scores will increase and decrease over time, while overall consistently increasing; this is indicative of disease-worsening.

While its weakness lies in known issues with consistency (inter- and intra-rater reliability) and sensitivity to change, it is relatively straight-forward to employ (D. S. Goodin et al., 2016; Meyer-Moock et al., 2014). Additionally, because it is commonly used, comparison of published studies is efficiently performed.

**Statistical properties and considerations** As described earlier, the EDSS consists of ordinal values ranging from 0.0 (no disability) to 10 (death), by increments of 0.5; higher scores indicate greater disability and poorer mobility. It has been shown to have a bimodal frequency distribution (J. Hutchinson & Hutchinson, 1995; Koziol, Frutos, Sipe, Romine, & Beutler, 1996; Sharrack, Hughes, C, Soudain, & Dunn, 1999; Willoughby & Paty, 1988), with peaks at 3.0 and 6.0 (Amato & Ponziani, 1999; Willoughby & Paty, 1988). It is also non-linear; this has been demonstrated as it relates to quality of life (Twork et al., 2010; Vickrey, Hays, Harooni, Myers, & Ellison, 1995). The EDSS has been demonstrated to be curvilinear relative to actual daily function (Cohen, Kessler, & Fischer, 1993). Because the EDSS are composite scores, the interval between scores are not the same; that is, the distance between scores do not have the same meaning (Meyer-Moock et al., 2014). Since change scores of a

particular value do not have consistent meaning, the starting position related to the change score is essential to interpreting the change score. Relatedly, use of a change score in this fashion requires controlling with a baseline EDSS measure. As mentioned previously, EDSS regions can be defined by the following classifications: 0-3.5 represent impairment based on functional system scores; 4.0-7.0 represent a combination of impairment and disability; 7.5 and greater represent requiring assistance for self-care (Amato & Ponziani, 1999). These distinctions are advantageous for collapsing scores for practical and modeling purposes.

The non-linear nature of EDSS eliminates several commonly used analytic methodologies; some authors have even argued against its use in the parametric setting at all (Amato & Ponziani, 1999). Weinshenker and colleagues showed that the “mean staying times” varies at each score; specifically, patients tend to remain at scores in the extreme (upper and lower) regions of the scale, more so than in the middle (at scores 3.0 through 5.0) (Amato & Ponziani, 1999; Weinshenker et al., 1989). These considerations are compelling reasons to investigate other, potentially more appropriate paths, by treating it ordinally or nominally, or even collapsing scores to minimize groups, and even inclusion of covariates to predict EDSS scores (or change scores). Therefore, should the appropriate model assumptions be met, ordinal, nominal and logistic regression are all appropriate, as is using a longitudinal model, should data collection occur for multiple time-points. The fact that an EDSS scores and change scores depend on the previous score suggests the appropriateness of utilizing Markovian methodology. In further support is the fact that the literature demonstrates a variety of methods of grouping those scores for

analysis, lending itself for investigation regarding the nature of grouping of EDSS scores as an analytic outcome.

## 2. The Nature of EDSS scores related to Markov models

The EDSS is well-suited to Markov modeling. The simplest Markov model is a Markov chain; in this sense, we would assume that the EDSS scores only depend on the most recent score, thus meeting the Markov property. However, this may not be conceptually reasonable, since various factors influence relapse and remission in MS. Therefore, utilizing logistic regression (nominal or ordinal) is appropriate, allowing for other factors to account for the probability of movement between scores. A Markov chain can be continuous or discrete; a discrete chain is one whose states and associated probabilities are measured over fixed time points. A continuous chain is one whose states do not occur at specific intervals. While disability may change at varying intervals, it is generally measured at semi-annual clinician visits; this is true whether or not a patient is enrolled in a clinical trial. Therefore, using a discrete chain is certainly appropriate to the typical EDSS data collection structure.

While the 20 states of the EDSS can be modeled, appropriate grouping of states will also be explored. This concept of collapsing states in the Markov context is known as lumping, and incorporates evaluating of maintaining the Markov property after state-grouping has occurred. Further discussion of this topic is explored in Chapter

II (Theoretical Background). Grouping the EDSS scores would be of interest, considering the distributional and statistical properties described earlier, as related to mean staying time at varying scores, and the range of scores defined by specific dysfunction.

Because MS is a progressive disease characterized by relapse and remission, it may be reasonable that a more complex Markov model is appropriate. That is, a higher order model where a future EDSS score is not solely predicted based on current EDSS score, but also the one(s) recorded previously.

### 3. PDDS as a primary endpoint and Markov applicability

**Use, meaning, strengths, weaknesses** The PDDS is a patient-reported assessment tool that is used for NARCOMS and in other studies where patient-reported outcomes are of interest, where it is not practical or possible to apply the EDSS, or it is the only means of assessing disability and disease progression (such as in a self-report observational setting, like NARCOMS) (Coyle et al., 2017; Learmonth et al., 2013; Rizzo et al., 2004). It evaluates the areas of “mobility, hand function, vision, fatigue, cognition, bladder/bowel, sensory and spasticity” (Rizzo et al., 2004). Similar to the EDSS, the PDDS scores can be grouped according to what it measures. The strength of correlation between the EDSS and the PDDS has been demonstrated to be consistent regardless of disease severity (Learmonth et al., 2013). Like the EDSS, the PDDS is driven, overall, by the mobility and motor skills of the patient

(Learmonth et al., 2013). Relatedly, it is also influenced by a patient's perception of disability (Schwartz, Vollmer, & Lee, 1999).

**Statistical properties and considerations** The PDDS, like the DS it was based upon, is an ordinal scale ranging from 0 to 8 with the following indications: 0: normal mobility, 1: mild disability, 2: moderate disability, 3: gait disability, 4: early cane use, 5: late cane use, 6: bilateral support, 7: wheelchair/scooter, and 8: bedridden (Rizzo et al., 2004). While there is not a one-to-one correspondence between scores, there is a similarity in the symptoms they represent, and the PDDS can be utilized as a surrogate of the EDSS (Table 1) (Marrie, Cutter, Tyry, Vollmer, & Campagnolo, 2006).

**Table 1:** Correspondence between PDDS and EDSS disability scales

<b>Symptomatic representation</b>	<b>PDDS Score<sup>1</sup></b>	<b>EDSS Score <sup>2</sup></b>
Normal mobility, mild symptoms	0	0
Gait disability, no assistive devices <sup>3</sup>	3	4—4.5
Assistive device required	4, 5, 6	6—6.5
Wheelchair-bound	7	7
Bedridden	8	8

<sup>1</sup>Patient-Determined Disease Steps

<sup>2</sup>Expanded Disability Status Scale

<sup>3</sup>EDSS scores 4-5.5 describes patients with Gait Disability but without assistive devices

#### 4. The Nature of PDDS scores related to Markov models

Like the EDSS, the PDDS is also well-suited to Markov modeling; this is a logical extension of their demonstrated similarity. The PDDS can be modeled as a discrete chain, since it is measured at fixed (bi-annual, 6-month) intervals. While the 9 states



can be modeled, appropriate groupings (lumping) of states (PDDS scores) merits exploration. Grouping the PDDS scores is of interest, as doing so might provide more information to patients and clinicians regarding mobility status and disability progression.

#### E. Gaps and Contributions

Application of Markov processes to MS outcomes has been accompanied with limitations related to sample size, study duration, covariate selection, MS definition and diagnoses changes, and logic behind collapsing of mobility states. C. Wolfson and Confavreux (1987) created a Markov model to study the natural history of MS using a survival model to produce hazard ratios to represent the movement between disease states (1987; C. Wolfson & Confavreux, 1985). This was prior to the introduction of the PDDS and EDSS diagnostic tools and is based on data collected between 1956 and 1976, with a moderate sample size (N=278). Albert (1994) employed a five-state Markov model to analyze relapsing-remitting behavior based on his categorization of disease worsening; however, this study was performed in mice with experimental allergic encephalomyelitis (EAE). It occurred over 40 days in a small sample (N=10). Mandel and Betensky (2008) illustrated their Markovian time-to event estimation in an MS population, but with a moderate sample size of 267 MS patients. Covariate selection was not explained, beyond availability. Most recently, Healy and colleagues have published work using EDSS (and groups of EDSS scores) as outcomes in Markov chains to determine appropriate predictors

associated with change in disability level (Engler, Chitnis, & Healy, 2017; Healy & Engler, 2009). However, these papers do not demonstrate evaluation of the resulting chains for retention of the Markov property.

This work will yield contributions of statistical and public health importance: we shall develop a goodness of fit (GOF) test that will compare lumped and unlumped (original) matrices, to be used after employing the test for lumpability from Baran (2001) and Jernigan and Baran (2003) (further discussion is found in Chapter II, Section C). Additionally, this work will endeavor to fill some gaps in the nature of studies investigating Markov models to predict the EDSS score. Specifically, our work will focus on a comprehensive dataset, representative of RRMS patients from the CombiRx trial; and on a wider-range of MS patients from the NARCOMS registry. Both of these datasets have been extensively published and peer-reviewed.

1. This work will involve larger sample sizes than seen in most prior studies, namely that of the CombiRx dataset (N=1008) and the NARCOMS dataset (N=2047). It will utilize necessarily varying ranges of sample sizes for simulation purposes. A particular strength of the CombiRx dataset is its low rate missing EDSS observations (the lowest is 22% missing during the 36-month study period) due to rigorous follow-up procedures. The only other comparable study was published in 2013, with approximately 2400 participants (Mandel, Mercier, Eckert, Chin, & Betensky). The strength of the NARCOMS data are its long-term observation periods and wide-range of MS disease severity included.

2. The larger sample sizes in these existing datasets will allow us to more accurately generalize results to RRMS (and lower disability scores) and a wider range of MS disease severities (and higher disability scores), and draw conclusions.
3. Covariates for the model will be selected based on known conventional covariate-adjustments (such as age and gender). This is particularly important, as the number of predictors and the predictors, themselves, can heavily influence model fit, results and conclusions.
4. Any grouping performed on the outcome (EDSS, PDDS) will be done based on scientific importance and based on known distributional properties, not data availability.
5. Application of Markov methodology to the PDDS, and to the NARCOMS dataset, will be novel.
6. Evaluation of the retention of Markov property will be novel for NARCOMS and CombiRx datasets and for both outcomes (PDDS and EDSS).

## ***II. Theoretical background***

### **A. Introduction and summary, notation**

#### **1. Summary of stochastic processes and introduction of notation**

A stochastic process, represented by  $Y(t)$ , is a group of random variables “...defined on a common probability space...” indexed by a set,  $T$ , such that  $\{Y(t), t \in T\}$ , and states, represented by  $y(t)$ , are any values taken by the process (Bhat & Miller, 2002; Resnick, 2002). Very often and in the context of this work,  $T$  represents time. The state space contains the possible values that the states can take; states can be discrete (countable) or continuous. In a discrete state setting, the states are not required to be numerical, but can be categories. Likewise, the index set can be discrete or continuous; that is, if we consider the typical index of time,  $T$ , it can be discrete. The parameter space contains the range of values for the index, e.g.,  $t_1, t_2, \dots, t_n$  (Bhat & Miller, 2002). In the context of time, the subscripts of the index set represent the points in time a value was measured or collected.

#### **2. Summary of and introduction to Markov processes**

A Markov chain is a stochastic process (also known as a Markov process) by which a unit transitions from one state to another with a given probability, which follows the

conventional definitions and restrictions of probabilities (i.e.,  $0 \leq p \leq 1$  and  $\sum_i p_i = 1$ ). A Markov chain is one which follows the Markov property; that is, the current state is dependent only upon the most recent state, and no others. Bartlett (1950) described it as being a “...process the memory of which does not extend beyond the previous instant.” All Markov chains have a defined state space, denoted as  $S$ ; for the EDSS outcome, this implies a maximum of 20 states ( $s=20$ ). For the PDDS, this implies a maximum of 9 states ( $s=9$ ). They are often depicted in the form of square matrices which shows probabilities of movement between states. For instance, should we have the maximum 20 states, the transition matrix would have 400 cells in which to put probabilities of movement; the PDDS would have 81 cells. We will note here that many of these transition probabilities would be very small (approaching zero), due to the high improbability of achieving certain transitions; an example would be moving from an EDSS of 9.5 to 1.0 (where 1.0 is minimal disability observed in one functional system; and 9.5 is confined to bed and entirely dependent upon assistance for mobility). Later sections will describe the transition matrices in more detail.

The initial state probability vector ( $\mathbf{a}_k$ ) describes the probability distribution of the states at the beginning of the chain. The initial probability distribution is important for determining the probability of transitions from one state to another, say from  $i$  to  $j$  with probability  $p_{i,j}$  (Resnick, 2002). Concerns often arise over deciding what the initial states of the model should be; these considerations tend to be practical in nature, and less theoretical. The initial distribution can be determined based on

general population values, pilot data or baseline data. An initial-state probability distribution can be estimated (or defined) from the observed data; as has been performed in the literature, baseline values are utilized for this purpose (Bickenbach & Bode, 2001; Mandel et al., 2013). In context of MS clinical trials, it is reasonable to allow for the initial state to be that which with the patient begins, as recorded at baseline. In general, clinical trials will restrict the range of EDSS scores for inclusion (for example, CombiRx limited the EDSS to  $\leq 5.5$ ); therefore, for those states not captured due to inclusion criteria would be given an initial probability of zero. We can anticipate observing higher probabilities associated with these early states of EDSS, and probabilities of 0 for those higher than the inclusion criteria. This disease-specific issue will influence the initial distribution seen based on the inclusion criteria. Conversely, we can anticipate observing nonzero probabilities for a wider range of PDDS scores, as there were no inclusion criteria regarding disease severity in terms of PDDS scores.

## B. Description of Markov processes and chains

### 1. Properties of Markov chains

We will consider the situation of having discrete time points (or, indexing parameters) in the Markov chain for this work; the random variables  $Y(t_1), Y(t_2), \dots, Y(t_n)$  will have a specific kind of dependence among them, the simplest of which is a first-order dependence. A Markov chain may be finite or infinite, referring to the number of countable states (Bhat & Miller, 2002). Markov

chains allow us to estimate the probability of achieving a specific state (by way of transition probabilities) with any spatial separation (spacing) between these specific, finite states. Let the state space be denoted by  $S$ , such that  $S=[0,1,2...s]$ , where  $s$  is the maximum number of states in the chain. In order to build the chain, we must complete the following general steps.

(1) Identify an initial distribution of the initial states of the process. If the initial distribution is indicated by  $[a_k]$ , then  $\sum_{k=0}^s a_k = 1$  for  $k \geq 0$  where

$$0 \leq a_k \leq 1, \forall k$$

(2) Obtain transition probabilities associated with movement between states. Let these probabilities be denoted as  $P_{i,j}$ , where

a.  $0 \leq p_{i,j} \leq 1$  for every  $i, j = 0, 1, 2 \dots s$

b.  $i=0, 1, 2, \dots, s$  indicates the current state

c.  $j=0, 1, 2, \dots, s$  indicates the next state

d.  $\sum_{j=0}^s p_{i,j} = 1$ , therefore, each row of the matrix sums to 1.

(adapted from Resnick (2002)).

The probabilities associated with moving between these states are represented in a transition matrix, indicated by  $P_T$ ; a transition matrix is always square (with dimensions of  $s \times s$ ). Let the term,  $n$ -step, describe the gap of time between instance of the current step and the instance of the next step (Bhat & Miller, 2002). Then, if  $s = 3$  states, the transition matrix for any given time point,  $t$ , would take the form in Figure 1. The states indicated in the vertical vector on the outside left of the matrix represent the current states; the states indicated on the horizontal vector on the

outside top of the matrix represent the next state after the current one. The matrix entries are the probabilities of moving to the next state, given the current state.

**Figure 1** : Basic 3-state, one-step transition matrix for a given time,  $t$

---


$$P^T = \begin{matrix} & \begin{matrix} 0 & 1 & 2 \end{matrix} \\ \begin{matrix} 0 \\ 1 \\ 2 \end{matrix} & \begin{bmatrix} p_{0,0} & p_{0,1} & p_{0,2} \\ p_{1,0} & p_{1,1} & p_{1,2} \\ p_{2,0} & p_{2,1} & p_{2,2} \end{bmatrix} \end{matrix}$$


---

For example,  $p_{0,0}$  indicates probability of remaining in state 0; and  $p_{0,1}$  indicates probability of movement from state 0 to state 1, and so on, at any given time.

In building the chain, we will indicate it as  $[Y_t = i, t \geq 0]$ , where  $Y$  is expressed as the state  $i$  at time  $t$  (Ross, 2003). A Markov chain will meet the following properties:

(1)  $P(Y_{t=0} = k) = a_k$ , therefore we are establishing the initial probability distribution of the state space.

(2)  $P(Y_{t+1} = j | Y_t = i) = p_{i,j}$ , for all  $i, j \geq 0$ . This indicates each entry in the transition matrix is a conditional probability, and thus based on those rules and assumptions thereby associated with conditional probability. It is more typically expressed as:

$$P(Y_{t+1} = j | Y_t = i, Y_{t-1} = i_{t-1}, Y_{t-2} = i_{t-2}, \dots, Y_0 = i_0) = P(Y_{t+1} = j | Y_t = i) \\ = p_{i,j}$$

Property 2 describes the Markov property (or Markov dependence), as mentioned in the introduction (Bhat & Miller, 2002; Resnick, 2002). That is, the transition probability is conditioned only on the most recent state.



## 2. Higher order models

Models which depend only upon the current state and no other earlier states are called *first order*. Models of higher order describe situations where the probability of moving to the next state is conditioned on 2 or more previous states. For example, we will consider a Markov chain of order 2 (*second order*), such that  $\{Y_n, n = 0, 1, 2, 3 \dots\}$  and with 3 states,  $\{1, 2, 3\}$ . Then a second order chain can be described using the following transition matrix (Figure 2b), which is an extension of the basic 3-state matrix introduced via Figure 1 (Figure 2a). This is mathematically expressed by

$$\begin{aligned} P(Y_{t+1} = k | Y_t = i, Y_{t-1} = j_{t-1}, Y_{t-2} = j_{t-2}, \dots, Y_0 = j_0) \\ = P(Y_{t+1} = k | Y_t = i, Y_{t-1} = j) \\ = p_{i,j,k} \end{aligned}$$

Because it is not a square matrix, Figure 2b is not useful for many matrix manipulations; Bhat and Miller (2002) recommend putting the second order transition matrix in the format of a first order matrix for analytic convenience (Figure 2c). We can then write the third property of Markov chains as it relates to higher order chains as:

- (3) Extending property (2) from the previous section, we may state the following: if we condition on several previous states, the probability of interest,  $p_{i,j}$ , is the same, and a next probability has a dependence only on the current state while being independent of time  $t$ . Then for any chain,  $Y_n$ , integer,  $n$  ( $n = 0, 1, 2, 3 \dots$ ) and any state,  $s$  ( $s=0, 1, \dots, s$ ), we have the following conditional probability

$$P_{ij,k} = P(Y_n = k | Y_{n-1} = j, Y_{n-2} = i)$$

(adapted from Resnick (2002) and Bhat and Miller (2002)).

**Figure 2:** Second order, 3-state, transition matrix for a given time,  $t$

**2a:** First order, 3-state probability transition matrix

$$P^T = \begin{matrix} & \begin{matrix} 0 & 1 & 2 \end{matrix} \\ \begin{matrix} 0 \\ 1 \\ 2 \end{matrix} & \begin{bmatrix} p_{0,0} & p_{0,1} & p_{0,2} \\ p_{1,0} & p_{1,1} & p_{1,2} \\ p_{2,0} & p_{2,1} & p_{2,2} \end{bmatrix} \end{matrix}$$

**2b:** Second order, 3-state, probability transition matrix

$$P^T = \begin{matrix} & \begin{matrix} 0 & 1 & 2 \end{matrix} \\ \begin{matrix} 00 \\ 01 \\ 02 \\ 10 \\ 11 \\ 12 \\ 20 \\ 21 \\ 22 \end{matrix} & \begin{bmatrix} p_{00,0} & p_{00,1} & p_{00,2} \\ p_{01,0} & p_{01,1} & p_{01,2} \\ p_{02,0} & p_{02,1} & p_{02,2} \\ p_{10,0} & p_{10,1} & p_{10,2} \\ p_{11,0} & p_{11,1} & p_{11,2} \\ p_{12,0} & p_{12,1} & p_{12,2} \\ p_{20,0} & p_{20,1} & p_{20,2} \\ p_{21,0} & p_{21,1} & p_{21,2} \\ p_{22,0} & p_{22,1} & p_{22,2} \end{bmatrix} \end{matrix}$$

**2c:** Transition matrix useful for matrix operations; first-order form

$$\begin{matrix} & \begin{matrix} 00 & 01 & 02 & 10 & 11 & 12 & 20 & 21 & 22 \end{matrix} \\ \begin{matrix} 00 \\ 01 \\ 02 \\ 10 \\ 11 \\ 12 \\ 20 \\ 21 \\ 22 \end{matrix} & \begin{bmatrix} p_{00,0} & p_{00,1} & p_{00,2} & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & p_{01,0} & p_{01,1} & p_{01,2} & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & p_{02,0} & p_{02,1} & p_{02,2} \\ p_{10,0} & p_{10,1} & p_{10,2} & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & p_{11,0} & p_{11,1} & p_{11,2} & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & p_{12,0} & p_{12,1} & p_{12,2} \\ p_{20,0} & p_{20,1} & p_{20,2} & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & p_{21,0} & p_{21,1} & p_{21,2} & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & p_{22,0} & p_{22,1} & p_{22,2} \end{bmatrix} \end{matrix}$$

### 3. Treatment of time

To this point, the Markov concepts discussed have been under the consideration that the transition probabilities do not change with time. A homogenous Markov chain (also known as stationary) is one which has the same probability of achieving a

certain state, independent of time,  $t$ . This means the process is time-invariant. In general, most Markov processes are not homogenous in a real-world setting (Resnick, 2002). Considering the implications for MS, because of the progressive nature of EDSS and MS, it is not reasonable to assume this temporal homogeneity (stationarity) in the Markov chain; however, this property may be true under specific settings. A chi-squared test of stationarity exists to evaluate this matrix property. Pooled transition probabilities (over the entire time period in question) are placed in a single matrix; then, the data are divided into a specified number of time periods. Thus, it evaluates the null hypothesis that the transition probabilities are the same between the pooled and sub-sampled matrices, versus the alternative that the probabilities are different (Bickenbach & Bode, 2001). The test statistic has an asymptotic chi-squared distribution, with  $\sum_{i=1}^N (a_i - 1)(b_i - 1)$  degrees of freedom; and takes the following form

$$Q^T = \sum_{t=1}^t \sum_{i=1}^n \sum_{j \in B_i} n_i(t) \frac{(\hat{p}_{i,j}(t) - \hat{p}_{i,j})^2}{\hat{p}_{i,j}}$$

Where  $Q^T$  = the test statistic

$T$  = time point of the matrix

$N$  = number of estimated parameters (transition probabilities)

$\hat{p}_{ij}$  = probability of transition from state  $i$  to state  $j$  in pooled matrix (across all  $T$  time points)

$\hat{p}_{i,j}(t)$  = probability of transition from state  $i$  to state  $j$  the matrix for time  $T$

$B_i$  = All nonzero transitions for the entire sample

$A_i$  = Number of subsamples of  $T$  where nonzero subsamples are in the  $i^{\text{th}}$  row

It would be of interest to evaluate this for the disability measures in an effort to determine where this temporal heterogeneity in the disease process occurs.

#### 4. Classification of states

States are classified according to how they relate to other states in the chain. State  $j$  is accessible from state  $i$  if there is a nonzero probability of reaching state  $j$  from state  $i$ , at any time point  $t$ . Any two states that are accessible with one another are said to communicate with one another; those states that communicate are defined as being in the same class. If  $i$  and  $j$  communicate, then this is expressed as  $i \leftrightarrow j$ . States in communication have the following relationships:

- (1) A state communicates with itself (state  $i$  communicates with state  $i$ ).
- (2) If  $i$  communicates with  $j$ , then  $j$  communicates with  $i$ .
- (3) If  $i$  communicates with  $j$ , and  $j$  communicates with  $k$ , then  $i$  communicates with  $k$ .

(adapted from Ross (2003)).

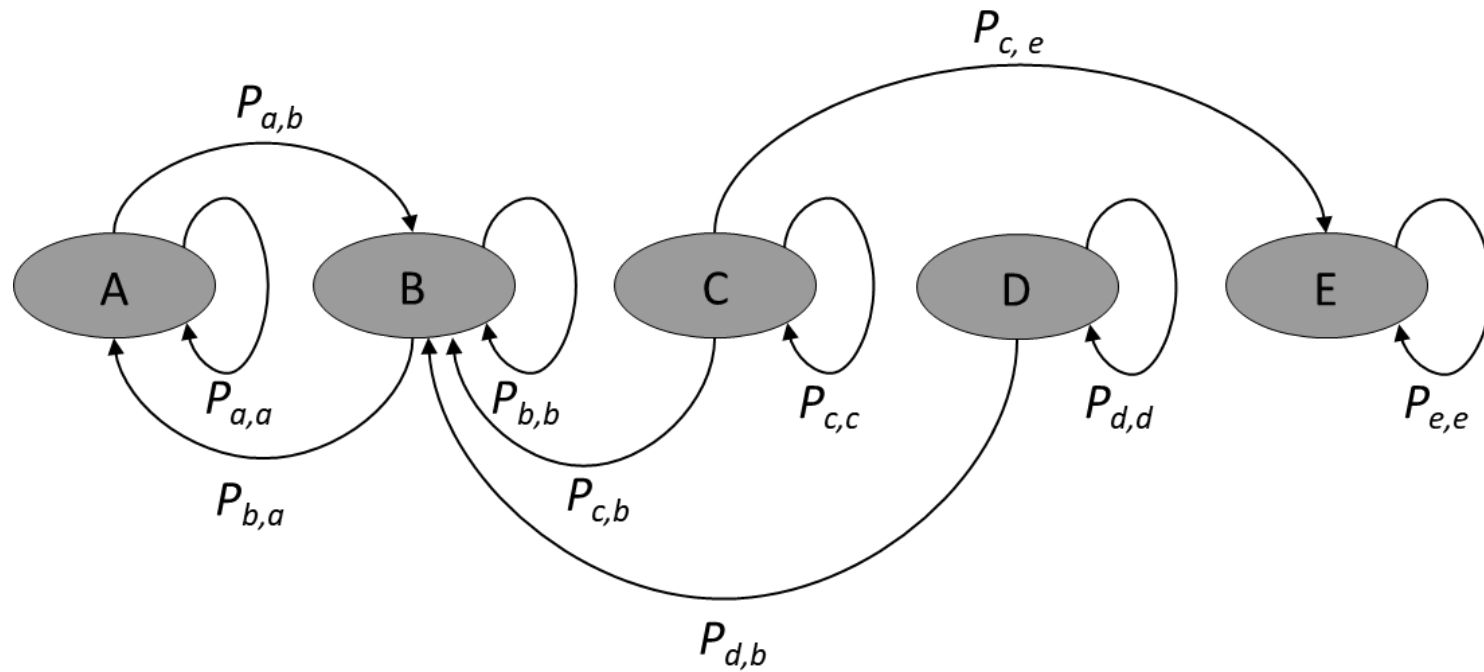
- 5. Classes are disjoint divisions in the state space; a Markov chain is *irreducible* if all states in the chain communicate with one another (therefore, there is a single class) (Bhat & Miller, 2002; Ross, 2003). Applying this to the example in Figures 3 and 4: there are 4 classes in this chain:  $\{A, B\}$ ,  $\{C\}$ ,  $\{D\}$  and  $\{E\}$ .

A state may be classified in the following ways; these states are most easily examined by creating a diagram of the chain from the transition matrix (Figures 3, 4).

- (1) Absorbing states are those, which, upon reaching them, there is zero-probability of leaving. That is, the transition probability of remaining in the state is 1, such that  $p_{i,i} = 1$  (state E, Figure 3).
- (2) Recurrent states are those that will be visited an infinite number of times and thus their probability of ever returning, is 1 (states A and B; Figure 3).
- (3) Transient states are those which, upon leaving, there is a possibility (nonzero probability) of never returning (state C; Figure 3).
- (4) There is a special case for transient states where, upon leaving, there is zero probability of returning to that state. We shall refer to such states as super-transient (state D; Figure 3).

**Figure 3:** Diagram of transitions between states A, B, C, D and E in a Markov Chain

---



---

Arrows/connecting lines above the states represent transition to higher states; those below represent transitions to lower states or remaining in the same state.

Therefore, depending on what is being studied, probabilities might represent a move to the next state, to a previous state, remain in the same state (recurrent) or reach a state and be unable to move out of it (absorbing) (Resnick, 2002; Ross, 2003). In the case of EDSS, we may consider this as probability of transitioning to a higher EDSS score, lower EDSS score and remaining with same EDSS score; the same considered transitions would be true for the PDDS.

**Figure 4:** 5-State transition matrix associated with the Markov chain depicted in Figure 3

---


$$P = \begin{matrix} & \begin{matrix} A & B & C & D & E \end{matrix} \\ \begin{matrix} A \\ B \\ C \\ D \\ E \end{matrix} & \begin{bmatrix} p_{a,a} & p_{a,b} & p_{a,c} & p_{a,d} & p_{a,e} \\ p_{b,a} & p_{b,b} & p_{b,c} & p_{b,d} & p_{b,e} \\ p_{c,a} & p_{c,b} & p_{c,c} & p_{c,d} & p_{c,e} \\ p_{d,a} & p_{d,b} & p_{d,c} & p_{d,d} & p_{d,e} \\ p_{e,a} & p_{e,b} & p_{e,c} & p_{e,d} & p_{e,e} \end{bmatrix} \end{matrix} = \begin{matrix} & \begin{matrix} A & B & C & D & E \end{matrix} \\ \begin{matrix} A \\ B \\ C \\ D \\ E \end{matrix} & \begin{bmatrix} p_{a,a} & p_{a,b} & 0 & 0 & 0 \\ p_{b,a} & p_{b,b} & p_{b,c} & 0 & 0 \\ 0 & p_{c,b} & p_{c,c} & 0 & p_{c,e} \\ 0 & p_{d,b} & 0 & p_{d,d} & 0 \\ 0 & 0 & 0 & 0 & 1 \end{bmatrix} \end{matrix}$$


---

## C. Lumpability: collapsing states to reduce the size of the transition matrix

### 1. Description of lumpability

In the case of categorical data analysis or categorical predictors for continuous outcome, it is often the practice to collapse multiple categories into fewer categories. This can be to accommodate small cell size; to meet the theoretical assumptions of a specific test; because such granulation of a variable is not necessary; or to simplify interpretation. In the specific case of EDSS, it is often the case to collapse the 20-category scale to 3 or 4 groups, resulting in categories that are easier to interpret and often with more similar sub-sample sizes.

In the context of Markov chains, grouping categories of an outcome (states) together is called “lumping.” Lumping is the process of grouping states in a chain, such that the Markov property is preserved in the lumped chain (Barr & Thomas, 1977).

Performing these groupings in the context of Markov models is not as straightforward as in other settings because not all chains have the property of being lumpable (Bhat & Miller, 2002). Therefore, partitioning a transition matrix to achieve some grouping of the states might result in a process which is no longer Markovian, because this might interfere with the nature of dependence between the current and next states (Bhat & Miller, 2002). Lumpability applies to specific partitions of the state space; and each time states are aggregated, lumpability must be evaluated (Bhat & Miller, 2002; Kemeny & Snell, 1960).

If a chain is truly lumpable, then those rows desired to be combined will have the same sum across the probabilities. For example, if we wished to move from a 3x3 matrix to a 2x2 matrix and thus we wish to combine the rows A and B, then we have the following:

**Figure 5 : Lumping matrices**

<i>Unlumped Matrix, 3 × 3</i>		<i>Equal probability sums</i>	<i>Lumped Matrix, 2 × 2</i>	
$P =$	A			
	B			
	C			
	A	B	C	
	$p_{a,a}$	$p_{a,b}$	$p_{a,c}$	
	$p_{b,a}$	$p_{b,b}$	$p_{b,c}$	
	$p_{c,a}$	$p_{c,b}$	$p_{c,c}$	
	$p_{a,a} + p_{a,b} = p_{b,a} + p_{b,b}$			
	AB	C		
	$p_{a,a} + p_{a,b}$	$p_{a,c} + p_{b,c}$		
	$p_{c,a} + p_{c,b}$	$p_{c,c}$		



If this criterion is met, then states A and B can be lumped together into a single state, resulting in the second (lumped) matrix. The resulting chain then retains the Markov property.

## 2. Evaluating lumpability

Thomas and Barr (1977) developed an approximate chi-squared test to evaluate whether or not a chain is lumpable; this test was adapted and improved upon in a dissertation by Baran (2001) and published later by Jernigan and Baran (2003). The test is based on the usual concepts of a chi-squared test, where the observed probabilities of the larger (original) transition matrix are considered in context with the expected probabilities for the collapsed (lumped, smaller) matrix. The purpose of the test is that it evaluates whether probabilities can be grouped together, based on the observed values and associated estimated probabilities, while preserving the Markov property.

The chi-squared test evaluates the null hypothesis that the chain is lumpable based on the proposed lumping scheme, versus the alternative hypothesis that the chain is not lumpable using the lumping scheme. Therefore, the test must be performed for each lumping scheme proposed.

Let " $o_{k,j}$ " represent observed transitions from state " $k$ " to lump " $j$ " and let " $i$ " represent each lump. Then " $n$ " refers to the counts associated with the rows from

the observed transition matrix, and “m” refers to the counts associated with the rows for the lumped transition matrix. Let “n” also represent the number of states in the observed matrix, such that the observed matrix has dimension  $n \times n$ ; then “m” represents the number of lumps (states) in the lumped matrix, such that the lumped matrix is  $m \times m$ .

The test takes the following form, using lump, row and observed counts:

$$X^2 = \sum_{k=1}^n \sum_{j=1}^m \frac{\left(o_{k,j} - \frac{n_{k \cdot} m_{i \cdot j}}{m_{i \cdot}}\right)^2}{\frac{n_{k \cdot} m_{i \cdot j}}{m_{i \cdot}}}$$

Thus, the test statistic takes the familiar form where the numerator is the observed minus the expected counts, divided by expected counts, where there are as much chi-squared values as cells in the lumped matrix (therefore  $m \times m$  chi-squared values) (Baran, 2001; Jernigan & Baran, 2003).

As an example, we shall again, consider the simple case of a  $3 \times 3$  being lumped to form a  $2 \times 2$  matrix from Figure 5. The matrices can be expressed using these terms as follows (Figures 6, 7). The degrees of freedom (DF) are calculated as  $DF = (m - 1)(n - m)$  (Baran, 2001; Jernigan & Baran, 2003).

---

**Figure 6:** Original example matrix expressed with counts,  $3 \times 3$  dimension

---

$$\begin{array}{c} \text{A} \\ \text{B} \\ \text{C} \end{array} \begin{bmatrix} n_{a,a} & n_{a,b} & n_{a,c} \\ n_{b,a} & n_{b,b} & n_{b,c} \\ n_{c,a} & n_{c,b} & n_{c,c} \end{bmatrix} \begin{array}{c} n_{A \cdot} \\ n_{B \cdot} \\ n_{C \cdot} \end{array}$$


---

**Figure 7:** Lumped matrix expressed with counts,  $2 \times 2$  dimension

---

<i>Lumped matrix, showing summation of states</i>			<i>Final lumped matrix</i>	
	$  \begin{array}{c}  \text{AB} \\  \text{C}  \end{array}  \begin{bmatrix}  n_{a,a} + n_{a,b} + n_{b,a} + n_{b,b} + n_{b,b} & n_{a,c} + n_{b,c} \\  n_{c,a} + n_{c,b} & n_{c,c}  \end{bmatrix}  $	$\Rightarrow$	$  \begin{array}{c}  \text{AB} \\  \text{C}  \end{array}  \begin{bmatrix}  m_{ab,ab} & m_{ab,c} \\  m_{c,ab} & m_{c,c}  \end{bmatrix}  $	$  \begin{array}{c}  m_{AB.} \\  m_{C.}  \end{array}  $

---

#### D. Markov (transition) models

##### 1. Using covariates to predict state transitions

Markov models are also known as transition models. Modeling an outcome using the previous state (or outcome) as a predictor, along with other covariates, results in a transition model. These states are those nominal or ordinal outcomes we wish to model, over some time,  $T$ . The simplest case of a Markov model is what we have previously seen, where no other predictors are utilized outside of the current state, to predict the next state. Now, we shall consider a Markov model where covariates (in addition to the current state) are utilized to predict the next state. Therefore, if we indicate the outcome as  $Y$ , then for each time,  $t$ , we have an outcome of  $Y_1, Y_2, \dots, Y_t$  (Agresti, 2007). The model generates the transition probabilities for the transition matrix; an appropriate logistic regression model (binary, nominal, ordinal) is fit for each time point. A transition matrix is estimated for each time point; the probabilities produced at each time point are then associated with the

matrices. Multiple time points means repeated observations per participant; because the models are fit per time point, each observation is considered independent because it is treated separately, as the time points are analyzed separately (Table 2) (Agresti, 2007).

**Table 2:** Methods of estimating transition probabilities in logistic models

Model	Model expression	Transition probability
Binary	$\text{logit}[P(Y_t = j Y_{t-1} = i)] = \alpha + \theta_{t-1}y_{t-1} + \sum_{l=1}^v \beta_l x_l$	$P(Y_t = j Y_{t-1} = i) = \frac{\exp(\alpha + \theta_{t-1}y_{t-1} + \sum_{l=1}^v \beta_l x_l)}{1 + \exp(\alpha + \theta_{t-1}y_{t-1} + \sum_{l=1}^v \beta_l x_l)}$
Nominal <sup>1</sup>	$\text{logit}\left[\frac{P(Y_t = j Y_{t-1} = i)}{P(Y_t = J)}\right] = \alpha_j + \theta_{t-1}y_{t-1} + \sum_{l=1}^v \beta_l x_l$	$P(Y_t = j Y_{t-1} = i) = \frac{\exp(\alpha_j + \theta_{t-1}y_{t-1} + \sum_{l=1}^v \beta_l x_l)}{1 + \exp(\alpha_j + \theta_{t-1}y_{t-1} + \sum_{l=1}^v \beta_l x_l)}$
Ordinal <sup>2</sup>	$\text{logit}[P(Y_t \leq j Y_{t-1} = i)] = \alpha_j + \theta_{t-1}y_{t-1} + \sum_{l=1}^v \beta_l x_l$	$P(Y_t \leq j Y_{t-1} = i) = \frac{\exp(\alpha_j + \theta_{t-1}y_{t-1} + \sum_{l=1}^v \beta_l x_l)}{1 + \exp(\alpha_j + \theta_{t-1}y_{t-1} + \sum_{l=1}^v \beta_l x_l)}$

<sup>1</sup> Where each outcome (state) has its own intercept and set of coefficients; 0=baseline category (state in the Markov chain)

<sup>2</sup> Meeting the proportional odds assumption

Where  $j$ =current state for the time period=1, 2, ...  $J - 1$

$i$ =previous state for the time period

$\alpha$  =intercept

$x_1, x_2, \dots, x_v$  are the covariates

$\beta$  represents the coefficient of the associated predictor

$\theta$  represents the coefficient of the previous state used as a predictor

## 2. Higher order models

The models discussed to this point are first order, meaning each outcome only depend on the outcome observed at the most recent time point, and is used thusly in the model as a predictor. Higher order models depend on a greater number of previous outcomes; for example, second order models utilize the previous two time points in the model as predictors. As described earlier, transition matrix of a second order Markov chain is built on that of the first order chain (Figure 2). The matrix clearly indicates the way in with the  $j^{th}$  state is conditional upon the previous two states. The associated logistic model would be conditioned on the previous 2 time points, and thus would both function as predictors in the model.

We can express the logistic models seen earlier to reflect this second time point (Table 2). A similar process would be employed if a third- or higher-order model was desired (Agresti, 2007). In the simplest case (binary logistic regression) we would express the model as follows

$$\text{logit}[P(Y_t = j)|Y_{t-1} = i, Y_{t-2} = h] = \alpha + \theta_{t-1}y_{t-1} + \theta_{t-2}y_{t-2} + \sum_{l=1}^v \beta_l x_l$$

Utilizing higher order models in MS can be useful, when considering the nature of the disease (or what is being modeled). For example, knowing the most recent 2 EDSS states may be informative regarding how quickly mobility is deteriorating. Because there are multiple paths to an EDSS state, knowing the patients' previous states (previous disability history) amounts to a general understanding of whether

a patient is in remission (7.0 to 5.0), slowly progressing (4.0 to 4.5) or quickly progressing (for instance, 4.0 to 7.0, which may not be realistic).

## E. Context of Markov models and multiple sclerosis

### 1. Distribution of EDSS

Study data will influence the initial distribution and overall distributions of EDSS.

First, the trial design, itself, will impact the EDSS distribution, simply based on inclusion/exclusion criteria. The initial distribution (the baseline scores) might mimic the distribution of EDSS in the general population or may be influenced by the study design. In the case of CombiRx, patients were excluded from enrollment if they had an EDSS of 6.0 or higher. Because patients are enrolled for the more mobile portion of their disease, we will lack information about higher EDSS scores (greater disability) during the earlier part of the study. Depending on the specific nature and course of each participant's disease, this will likely also extend to the later part of the study, where we will again lack information regarding higher order EDSS scores. Second, there is bias present towards the patient pool, itself, as there might be something unique about those who engage in clinical trials versus not.

## 2. Distribution of PDDS

Survey data will influence the initial distribution and overall distributions of PDDS. First, the collection method, itself, will impact the PDDS distribution, based on factors at enrollment and over time. The initial distribution (the baseline scores) might mimic the distribution of PDDS in the general population or may be influenced either by the self-reported nature of the variable and/or such factors as age, disease duration, or disease severity. Response-rate over time might be influenced by disease progression, experience of relapse, and other social or medical factors (i.e., changing marital status, depression, degenerating fine motor skills, requiring a caregiver to complete the survey). Because patients are engaged from enrollment onwards, and are encouraged to participate for as long as possible; therefore, there is potential for greater insight for later disease stages. Therefore, there is opportunity to learn more about higher PDDS scores the disease and patient experience. We will therefore be limited to the participants' continuity of response. Second, there is bias present within the participant pool, itself, as there might be something unique about those who participate in observational studies versus not.

## 3. Relevance of lumpability to EDSS and PDDS

The goal of the work is two-fold: application of an analytic method on an outcome for which previous methods are not entirely satisfactory, and using the EDSS in the most simple form possible, to facilitate interpretation in a clinical setting. The EDSS



is a complex outcome, and as mentioned in the introduction, it has 20 states, which would result in an ungainly 400-cell transition matrix. Grouping the EDSS scores in a meaningful way might provide insight to the disease and better reflect the disease, itself.

The PDDS might have more patients whose mobility is worse than is observed in clinical trials, since it is obtained from patients in a registry. Because of this it is more likely to represent MS population-level information and have less sparse data for more advanced disease.

Statistically speaking, grouping the EDSS and PDDS would enable more tractable calculations and would allow for preservation of degrees of freedom in the model-building setting. Additionally, previous work of Markov model application to EDSS analysis did not evaluate the lumpability property. Therefore, should we determine the most appropriate aggregation of EDSS scores, it is critical to know whether or not it is appropriate to proceed using Markov chain methodology (Kemeny & Snell, 1960).

#### F. Model selection and Goodness of Fit: previous methods, proposed considerations

##### 1. Model selection to estimate parameters (transition probabilities)

Multiple aspects must be considered for rigorous model selection. First, the nature of the outcome will inform the choice of model type (Table 2); this is related to determining the appropriate aggregation of EDSS, as described above. Once this has

been established, the usual methods for detecting model fit must be employed. Specific consideration must be given regarding model fit between orders of models (e.g., between Markov transition models of order one and order two) and incorporation of random effects.

## 2. Goodness of fit

Several variations of the likelihood ratio test (LRT) have been developed and employed to evaluate fit for different criteria. Most recently, Mandel and Betensky (2008) employed likelihood ratio tests to compare models with/without random effects. In a later paper, Mandel and colleagues (2013) evaluated their models by comparing the predictive performance of the models. Specifically, they examined the expected proportion of transitions versus the observed ones; they also examined the binomial confidence intervals about the proportions. The order of Markov models have been evaluated using a chi-squared GOF test, as well as using the AIC and BIC (Baran, 2001; Jimoh & Webster, 1996; Katz, 1981). While not a requirement for using these criteria, chains of different orders are truly nested models. We have already described the use of the chi-squared test of lumpability (Baran, 2001; Jernigan & Baran, 2003). The tests of stationarity and others utilize maximum likelihood methods and are describe in the work of Billingsley (1961).

We are interested in extending the concept of lumping to determine if, given a matrix is lumpable according to a given lumping scheme, it is the best fit to the data.

Avenues of investigation will include development of an LRT to compare lumped and unlumped matrices, as well as a measure of comparing two lumping schemes for a given original, unlumped chain.

### ***III. Foundational theoretical work and development of a Goodness of Fit Test***

#### **A. Introduction**

The chi-squared test of lumpability by Thomas and Barr (1977) was updated by Jernigan and Baran (2001; 2003). As an extension, it is of interest to consider the usefulness and appropriateness of lumping states at all. If, according to the chi-squared testing of lumpability the specified lumping scheme is acceptable and the chain is lumpable, a question naturally follows: does use of this scheme lead to the best fit for the data? In order to compare the fit of the lumped and unlumped matrices, we propose two goodness of fit tests using likelihood ratio and Pearson formulations, and explore the properties of these tests via simulation.

#### **B. Methods**

##### **1. Description of simulated data and states**

Data were simulated based upon matrices that were designed to be lumpable. These true matrices of transition probabilities,  $\mathbf{B}_0$  and row counts were  $4 \times 4$  (four states:  $A, B, C, D$ ), and whose lumped dimension was  $2 \times 2$  (two states:  $AB, CD$ ). Four matrices were considered and used to generate chains, which were later lumped using the same lumping scheme were developed to further assess our

proposed methodology; these matrices were numbered 1 through 4 (Table 3). All chains shared the same initial probability distribution, such that when frequencies were simulated, equal probabilities were assumed for each state for the initial distribution, such that  $\mathbf{S}_0 = [0.25 \ 0.25 \ 0.25 \ 0.25]$ .

**Table 3:** Transition probabilities used for simulation

Matrix	“True” transition probabilities, $\mathbf{B}_0$
1*	$\begin{bmatrix} 0.48 & 0.48 & 0.02 & 0.02 \\ 0.48 & 0.48 & 0.02 & 0.02 \\ 0.03125 & 0.03125 & 0.46875 & 0.46875 \\ 0.03125 & 0.03125 & 0.46875 & 0.46875 \end{bmatrix}$
2	$\begin{bmatrix} 0.44 & 0.46 & 0.04 & 0.06 \\ 0.4475 & 0.4275 & 0.0625 & 0.0625 \\ 0.03 & 0.02 & 0.47 & 0.48 \\ 0.03125 & 0.03125 & 0.46875 & 0.46875 \end{bmatrix}$
3	$\begin{bmatrix} 0.44 & 0.46 & 0.04 & 0.06 \\ 0.4475 & 0.4275 & 0.0625 & 0.0625 \\ 0.029 & 0.021 & 0.46 & 0.49 \\ 0.027 & 0.0555 & 0.47 & 0.4475 \end{bmatrix}$
4	$\begin{bmatrix} 0.48 & 0.48 & 0.02 & 0.02 \\ 0.4375 & 0.4375 & 0.0625 & 0.0625 \\ 0.03125 & 0.03125 & 0.46875 & 0.46875 \\ 0.03125 & 0.03125 & 0.46875 & 0.46875 \end{bmatrix}$

\*Designed to be perfectly lumpable

## 2. Evaluation of lumpability

We employed Thomas and Barr (1977) chi-squared test of lumpability to examine the proposed lumping scheme (Figure 8). The proposed four-state (unlumped) process is formally expressed as  $S = \{A, B, C, D\}$ . The proposed two-state lumped process is formally expressed as  $S = \{AB, CD\}$  (Figure 8). Therefore, the states A

and B and states C and D are lumped together to form a 2-state process from a 4-state process. Multiple sample sizes (number of transitions) will be considered, as described in the next section.

**Figure 8:** Lumping scheme implementation

Unlumped matrix ( $4 \times 4$ ) expressed as probabilities					Unlumped matrix expressed as counts						
	A	B	C	D		A	B	C	D		
A	$p_{a,a}$	$p_{a,b}$	$p_{a,c}$	$p_{a,d}$	$p_{A\cdot} = 1$	A	$n_{a,a}$	$n_{a,b}$	$n_{a,c}$	$n_{a,d}$	$n_{A\cdot}$
B	$p_{b,a}$	$p_{b,b}$	$p_{b,c}$	$p_{b,d}$	$p_{B\cdot} = 1$	B	$n_{b,a}$	$n_{b,b}$	$n_{b,c}$	$n_{b,d}$	$n_{B\cdot}$
C	$p_{c,a}$	$p_{c,b}$	$p_{c,c}$	$p_{c,d}$	$p_{C\cdot} = 1$	C	$n_{c,a}$	$n_{c,b}$	$n_{c,c}$	$n_{c,d}$	$n_{C\cdot}$
D	$p_{d,a}$	$p_{d,b}$	$p_{d,c}$	$p_{d,d}$	$p_{D\cdot} = 1$	D	$n_{d,a}$	$n_{d,b}$	$n_{d,c}$	$n_{d,d}$	$n_{D\cdot}$

Lumped matrix ( $2 \times 2$ ), showing summation of states		Final lumped matrix				
AB	CD	AB	CD			
AB	$n_{a,a} + n_{a,b} + n_{b,a} + n_{b,b}$	$n_{a,c} + n_{b,c} + n_{a,d} + n_{b,d}$	AB	$m_{ab,ab}$	$m_{ab,cd}$	$m_{AB\cdot}$
CD	$n_{c,a} + n_{c,b} + n_{d,a} + n_{d,b}$	$n_{c,c} + n_{c,d} + n_{d,c} + n_{d,d}$	CD	$m_{cd,ab}$	$m_{cd,cd}$	$m_{CD\cdot}$

- Proposed tests: exploration of the goodness of fit test statistics and degrees of freedom

The proposed GOF tests evaluate the same hypotheses: the null hypothesis that the lumped matrix is a better fit to the data, versus the alternative that the unlumped matrix is a better fit to the data. The first is a likelihood ratio test (LRT). The second proposed test is a Pearson test, with the traditional formulation of observed and expected components. Also developed is a measure of the degree of difference between the unlumped, perfectly lumpable matrix, and an actual matrix, the  $W$ -score. Both tests' distributional properties were explored. Degrees of freedom were

empirically determined via histogram with distributional overlay for the simulated test statistics.

### 3.1 Likelihood estimation

The first proposed GOF test employs a likelihood ratio; the LRT statistic takes on the usual form of the numerator being the likelihood of the data under the unlumped matrix, and the denominator being the likelihood under the lumped transition matrix. To that end, realization of the LRT statistic and its components (i.e., numerator and denominator of the test statistic) were visualized via histogram and summarized using means, medians, minima and maxima.

Let  $P$  describe the probability transition matrix; and let the maximum likelihood estimator (MLE) for a transition probability,  $p_{i,j}$ , be expressed as  $\hat{p}_{i,j} = \frac{n_{i,j}}{\sum_{j=1}^n n_{i,j}}$  (Bickenbach & Bode, 2001); the MLE for  $p_{i,j}$  is asymptotically normal and unbiased (Anderson & Goodman, 1957). Now the log-likelihood of the data given  $p$  is as

$$\ln(L(P)) = \sum_{i,j=1}^n n_{i,j} \ln(p_{i,j})$$

where  $n_{i,j}$  represents the number of observed transitions from  $i$  to  $j$  and  $p_{i,j}$  represents the transition probability between states  $i$  and  $j$ .

Now, let  $p_{i,j}$  and  $\hat{p}_{i,j}$  represent the associated transition probabilities for the unlumped (most often, the observed) matrix. Let the likelihood be expressed as

$$\ln(L(\mathbf{P})) = \sum_{i=1}^n \sum_{j=1}^n n_{i,j} \ln(p_{i,j})$$

(Bhat & Miller, 2002).

Now, let  $q_{i,j}$  and  $\hat{q}_{i,j}$  represent the associated transition probabilities for the lumped matrix; let  $m_{i,j}$  represent the observed transitions between the lumped states. Then the LRT statistic which compares the unlumped ( $\mathbf{P}$ ) and lumped ( $\mathbf{Q}$ ) matrices takes the form:

$$\begin{aligned} -2 \ln(L) &= 2[L(\mathbf{Q}) - L(\mathbf{P})] \\ &= 2 \left[ \sum_{i=1}^m \sum_{j=1}^m m_{i,j} \ln(q_{i,j}) - \sum_{i=1}^n \sum_{j=1}^n n_{i,j} \ln(p_{i,j}) \right] \end{aligned}$$

### 3.2 Pearson test formulation

The Pearson goodness of fit test statistic,  $X_p^2$ , will take the conventional form using observed and expected values, based on the unlumped matrix. Again, let  $n_{ij}$  represents the number of observed transitions from  $i$  to  $j$ . If any zero-cells exist in the observed transition matrix, the usual 0.5-integer correction is used, such that the “0” is replaced with “0.5.” The expected number of transitions are expressed as  $p_{i,j}n_{i,j}$ . The values that contribute to the Pearson test statistic are in the context of the unlumped matrix.

$$X_p^2 = \sum \frac{(n_{i,j} - p_{i,j}n_{i,j})^2}{p_{i,j}n_{i,j}}$$



### 3.3 $W$ -score

In an effort to understand how lumpable a matrix is compared to a perfectly lumpable matrix, we developed a statistic which we denote as  $W$ . This describes a property of the unobserved, perfectly lumpable transition matrix and is a measure of the magnified of difference between a perfectly lumpable and actual matrix.  $W$  takes into account the observed number of states, the number of observed states contributing to each lump and the probability of any state contributing to each lump. A perfectly lumpable matrix has a  $W$  of 0. Larger values of  $W$  indicate the chain is further from a perfectly lumpable matrix; and thus, it should be more likely that we reject the null hypothesis lumpability in favor of the unlumped matrix. Let  $\bar{p}_l$  represent the mean transition probability for a lump,  $l$ , under the proposed lumping scheme. Let  $n_l$  indicate the number of cells from the unlumped matrix contributing to that lump.

$$W = \sum_{j \in l} \frac{(p_{i,j} - \bar{p}_l)^2}{n_l}$$

When performing the simulations, this statistic is useful to measure how lumpable the starting matrix was so that we could anticipate the results of the test of lumpability.

### 3.4 Simulation methodology

In order to determine the distributional properties of the proposed test statistics, we simulated data based on 4 underlying transitions matrices, each  $4 \times 4$  in dimension, using the reported initial distribution (Table 3). These simulation steps are summarized in Table 4. We considered 20 sample sizes to explore the LRT and Pearson statistics; these sample sizes ranged from 50 to 1,000, in increments of 50. For each of these sample sizes 1,000 iterations were performed, generating 1,000 observed transition matrices. All methods remain the same for each sample size, iteration and underlying Markov matrix. Only sample sizes of 50, 200, 1000 and 5000 were examined for distributional investigations.

First, the transitions in each row were simulated based on the fact that each row follows a multinomial distribution, based on the 4 probabilities of the 4-state chain ( $p_i$ , where  $i = 1, 2, \dots, 4$ ) for each of the 4 possible transitions and thus  $n(n - 1)$  parameters (transition probabilities) were estimated based on the observed chain. Second, lumped matrix was then created, by grouping states based on the proposed lumping scheme. Third, each matrix pair (observed, unlumped and associated lumped) was then evaluated with the chi-squared test of lumpability (Jernigan & Baran, 2003; Thomas & Barr, 1977). This test is considered Stage 1. Fourth and finally, if the chain passed Stage 1 such that the chain was considered lumpable, it was then evaluated under Stage 2 whereby it was evaluated for goodness of fit, using the proposed GOF tests. Thus, LRTs and Pearson statistics were then calculated for each pair of matrices.

The distributions for the LR and Pearson tests were then visualized via histogram, from which the degrees of freedom were empirically demonstrated and further examined via the Kolmogorov-Smirnov (KS) and Anderson-Darling (AD) tests. We used the KS and AD tests to compare empirical distribution of the test statistics to the Chi-square distribution.

**Table 4 :** Summary of general simulation steps

---

1. Simulate initial states using $S_0$
2. <i>Simulation</i> : simulate observed frequency transition matrix based on actual probability transition matrix, $B_0$ <ol style="list-style-type: none"> <li>Matrix is dimension <math>n \times n</math> and sample size <math>N</math>, based on the observed <math>S_0</math>.</li> <li>Each row is simulated separately using the multinomial distribution.</li> </ol>
3. <i>Lumping</i> : aggregate the frequency transition matrix of $n$ states to $m$ states <ol style="list-style-type: none"> <li>Done according to a pre-specified lumping scheme.</li> <li>Lumped matrix is <math>m \times m</math> in dimension.</li> </ol>
4. <i>Stage 1 testing</i> : evaluate lumpability by performing chi-square test of lumpability
5. <i>Stage 2 testing</i> : evaluate goodness of fit of lumping scheme by performing <ol style="list-style-type: none"> <li>Likelihood ratio test and Pearson test (half-integer corrected)</li> </ol>
6. Iterate Steps 0-4 1000 times for sample sizes 50 to 1000 by 50.

---

#### 4. Linear Algebra

##### 4.1 Implementation of a lumping scheme and performing the test of lumpability

Mathematically, the simulation for a single iteration goes as follows, using the previously defined initial distribution,  $S_0$ , and the true transition matrix for each chain. Matrix algebra and notation for lumping has been adapted and extended from

Baran (2001). We shall be using the conventional symbol of the Hadamard product, “ $\circ$ ”, indicating elementwise multiplication (Anton, 2010; Styan, 1973). To account for elementwise division, we shall use the Hadamard division operator,  $\oslash$ .

For simplicity, we will outline the linear algebra only for Chain 3, such that its true transition matrix is defined as  $B_0$ , such that

$$B_0 = \begin{bmatrix} 0.44 & 0.46 & 0.04 & 0.06 \\ 0.4475 & 0.4275 & 0.0625 & 0.0625 \\ 0.029 & 0.021 & 0.46 & 0.49 \\ 0.027 & 0.0555 & 0.47 & 0.4475 \end{bmatrix}$$

After simulation is performed, the resulting frequency matrix is defined as follows,

$$N = \begin{bmatrix} n_{a,a} & n_{a,b} & n_{a,c} & n_{a,d} \\ n_{b,a} & n_{b,b} & n_{b,c} & n_{b,d} \\ n_{c,a} & n_{c,b} & n_{c,c} & n_{c,d} \\ n_{d,a} & n_{d,b} & n_{d,c} & n_{d,c} \end{bmatrix}$$

The initial state distribution,  $\mathbf{S}_0 = [0.25 \ 0.25 \ 0.25 \ 0.25]$ , established approximate equal proportions from each starting state for given time point. We simulated from 400 total transitions, indicating we started with approximately 100 transitions per row in the entire matrix at a single time point.

Then an example of simulated transitions and estimated probability transition matrices are

$$N = \begin{bmatrix} 44 & 46 & 5 & 6 \\ 45 & 43 & 6 & 5 \\ 3 & 2 & 46 & 47 \\ 3 & 6 & 47 & 46 \end{bmatrix}$$

$$N_{pr} = \begin{bmatrix} 0.4356436 & 0.4554455 & 0.049505 & 0.0594059 \\ 0.4545455 & 0.4343434 & 0.0606061 & 0.0505051 \\ 0.0306122 & 0.0204082 & 0.4693878 & 0.4795918 \\ 0.0294118 & 0.0588235 & 0.4607843 & 0.4509804 \end{bmatrix}$$

The lumping matrix,  $K$ , has dimension  $m \times n$ . Entries of “1” represent inclusion in the lumped state; correspondingly, entries of “0” indicate a state(s) are not include in the lumped state. Therefore, as we are interested in combining states  $A$  and  $B$  in a single state,  $AB$ ; and in combining states  $C$  and  $D$  into a single state,  $CD$ , the appropriate lumping matrix is

$$K = \begin{bmatrix} 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 \end{bmatrix}$$

The intermediate matrix of state-to-lump transitions is defined as  $V$ , with dimension  $n \times m$ .

$$\begin{aligned} V &= N \cdot K^T \\ &= \begin{bmatrix} 44 & 46 & 5 & 6 \\ 45 & 43 & 6 & 5 \\ 3 & 2 & 46 & 47 \\ 3 & 6 & 47 & 46 \end{bmatrix} \cdot \begin{bmatrix} 1 & 0 \\ 1 & 0 \\ 0 & 1 \\ 0 & 1 \end{bmatrix} \\ &= \begin{bmatrix} 90 & 11 \\ 88 & 11 \\ 5 & 93 \\ 9 & 93 \end{bmatrix} \end{aligned}$$

Next, the lumped matrix,  $M$ , is calculated as follows, with dimension  $m \times m$ .

$$\begin{aligned} M &= K \cdot N \cdot K^T \\ &= \begin{bmatrix} 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 \end{bmatrix} \cdot \begin{bmatrix} 44 & 46 & 5 & 6 \\ 45 & 43 & 6 & 5 \\ 3 & 2 & 46 & 47 \\ 3 & 6 & 47 & 46 \end{bmatrix} \cdot \begin{bmatrix} 1 & 0 \\ 1 & 0 \\ 0 & 1 \\ 0 & 1 \end{bmatrix} \\ &= \begin{bmatrix} 178 & 22 \\ 14 & 166 \end{bmatrix} \end{aligned}$$

We shall define the row sums of the observed frequency matrix as a column vector,  $n_k$ , of dimension  $n \times 1$ . Similarly, the row sums of the lumped frequency matrix is the column vector,  $m_i$ , of dimension  $m \times 1$ . In the context of our example, these are expressed as

$$n_k = \begin{bmatrix} 101 \\ 99 \\ 98 \\ 102 \end{bmatrix} \quad \text{and} \quad m_i = \begin{bmatrix} 200 \\ 200 \end{bmatrix}$$

The following 4 matrices are defined for the purposes of expressing the linear algebra to reflect the proposed lumping scheme.  $N_{i+}$  is the matrix of observed row sums, where the vector is repeated per number of desired lumps. For instance, if a two-state chain is desired, then two-lumps is the desired result, and so the column will be repeated twice; thus,  $N_{i+}$  has dimension  $n \times m$ .

$$N_{i+} = \begin{bmatrix} 101 & 101 \\ 99 & 99 \\ 98 & 98 \\ 102 & 102 \end{bmatrix}$$

$M_{sx}$  is the matrix of repeated rows of the lumped matrix, as necessary.

$$\begin{aligned} M_{sx} &= K^T \cdot M \\ &= \begin{bmatrix} 1 & 0 \\ 1 & 0 \\ 0 & 1 \\ 0 & 1 \end{bmatrix} \cdot \begin{bmatrix} 178 & 22 \\ 14 & 166 \end{bmatrix} \\ &= \begin{bmatrix} 178 & 22 \\ 178 & 22 \\ 14 & 186 \\ 14 & 186 \end{bmatrix} \end{aligned}$$

$M_{s+}$  is the matrix of repeated columns of  $m_i$ , reflecting the number of desired lumps.

The number of columns is the number of lumps.

$$M_{s+} = \begin{bmatrix} 200 & 200 \\ 200 & 200 \\ 200 & 200 \\ 200 & 200 \end{bmatrix}$$

If considering the conventional representation of a Chi-square test statistic formula where the numerator is the square of observed minus expected, then the square-root of this is expressed as follows:

$$\begin{aligned} Q &= V - N_{i+} \circ (M_{sx} \oslash M_{s+}) \\ &= \begin{bmatrix} 90 & 11 \\ 88 & 11 \\ 5 & 93 \\ 9 & 93 \end{bmatrix} - \begin{bmatrix} 101 & 101 \\ 99 & 99 \\ 98 & 98 \\ 102 & 102 \end{bmatrix} \circ \left( \begin{bmatrix} 178 & 22 \\ 178 & 22 \\ 14 & 186 \\ 14 & 186 \end{bmatrix} \oslash \begin{bmatrix} 200 & 200 \\ 200 & 200 \\ 200 & 200 \\ 200 & 200 \end{bmatrix} \right) \\ &= \begin{bmatrix} 0.11 & -0.11 \\ -0.11 & 0.11 \\ -1.86 & 1.86 \\ 1.86 & -1.86 \end{bmatrix} \end{aligned}$$

Then the Chi-square test of lumpability numerator,  $X_{Num}$ , is expressed as

$$\begin{aligned} X_{Num} &= Q \circ Q \\ &= \begin{bmatrix} 0.11 & -0.11 \\ -0.11 & 0.11 \\ -1.86 & 1.86 \\ 1.86 & -1.86 \end{bmatrix} \circ \begin{bmatrix} 0.11 & -0.11 \\ -0.11 & 0.11 \\ -1.86 & 1.86 \\ 1.86 & -1.86 \end{bmatrix} \\ &= \begin{bmatrix} 0.0121 & 0.0121 \\ 0.0121 & 0.0121 \\ 3.4596 & 3.4596 \\ 3.4596 & 3.4596 \end{bmatrix} \end{aligned}$$

The denominator,  $X_{Den}$ , is expressed as follows. Note that for ease of computation, at this stage, any zero-cells are indicated as missing values, and as such do not contribute to the test statistic.

$$\begin{aligned}
X_{Den} &= N_{i+} \circ (M_{sx} \oslash M_{s+}) \\
&= \begin{bmatrix} 101 & 101 \\ 99 & 99 \\ 98 & 98 \\ 102 & 102 \end{bmatrix} \circ \left( \begin{bmatrix} 178 & 22 \\ 178 & 22 \\ 14 & 186 \\ 14 & 186 \end{bmatrix} \oslash \begin{bmatrix} 200 & 200 \\ 200 & 200 \\ 200 & 200 \\ 200 & 200 \end{bmatrix} \right) \\
&= \begin{bmatrix} 89.89 & 11.11 \\ 88.11 & 10.89 \\ 6.86 & 91.14 \\ 7.14 & 94.86 \end{bmatrix}
\end{aligned}$$

In the usual calculations of the Chi-square test where two matrices of identical dimension are being tested (the observed and expected), there are the same number of component Chi-square statistics as there are cells in either matrix; these are then summed to produce the final overall test-statistic. In the case of the test of lumpability, the matrix of individual Chi-square components has  $n \times m$  dimension; thus, the product of the observed states and desired lumps is the number of individual values which will be summed to produce the Chi-square test statistic for the test of lumpability. For instance, in our particular example of moving from an observed 4-state chain to a 2-state lumped chain, we would have  $n \times m = 4 \times 2 = 8$  individual components, which would be summed to produce the overall test statistic evaluating lumpability of the chain according to the proposed lumping scheme. Accordingly, the individual Chi-square test statistic values representing the move from one state to one lump is obtained as

$$\begin{aligned}
X^2 &= (Q \circ Q) \oslash (N_{i+} \circ (M_{sx} \oslash M_{s+})) \\
&\equiv X_{Num} \oslash X_{Den}
\end{aligned}$$



Therefore,

$$\begin{aligned}
X^2 &= X_{Num} \oslash X_{Den} \\
&= \begin{bmatrix} 0.0121 & 0.0121 \\ 0.0121 & 0.0121 \\ 3.4596 & 3.4596 \\ 3.4596 & 3.4596 \end{bmatrix} \oslash \begin{bmatrix} 89.89 & 11.11 \\ 88.11 & 10.89 \\ 6.86 & 91.14 \\ 7.14 & 94.86 \end{bmatrix} \\
&= \begin{bmatrix} 0.0001346 & 0.0010891 \\ 0.0001373 & 0.0011111 \\ 0.5043149 & 0.0379592 \\ 0.4845378 & 0.0364706 \end{bmatrix}
\end{aligned}$$

This matrix results in a grand sum of  $X_s^2 = 1.0657546 \approx 1.07$ .

Baran (2001) updated the degrees of freedom for the Chi-square test of lumpability; this revision accounts for the presence of zero-cells in the matrices (Jernigan & Baran, 2003). Let  $M_s$  be the vector of the number of non-zero cells in each row of the lumped matrix,  $M$ . Then this vector,  $M_s$ , is  $m \times 1$  in dimension. Let the vector  $N_s$  represent the row sums of the lumping matrix,  $K$ ; then this vector is also  $m \times 1$  in dimension. The degrees of freedom of the test is the grand sum of the following matrix product:

$$\begin{aligned}
DF_z &= (M_s - 1) \circ (N_s - 1) \\
&= \left( \begin{bmatrix} 2 \\ 2 \end{bmatrix} - \begin{bmatrix} 1 \\ 1 \end{bmatrix} \right) \circ \left( \begin{bmatrix} 2 \\ 2 \end{bmatrix} - \begin{bmatrix} 1 \\ 1 \end{bmatrix} \right) \\
&= \begin{bmatrix} 1 \\ 1 \end{bmatrix} \circ \begin{bmatrix} 1 \\ 1 \end{bmatrix} \\
&= \begin{bmatrix} 1 \\ 1 \end{bmatrix}
\end{aligned}$$

This matrix results in a grand sum of 2, for 2 degrees of freedom. Based on these particular results, we have evidence to support the proposed lumping scheme for

the data, as we fail to reject the null hypothesis of lumpability at the  $\alpha=0.05$  level ( $p = 0.5869138 \approx 0.5869$ ).

It is important to note here the possibility of obtaining  $DF_z = 0$ ; in this case, the test of lumpability is not valid and the proposed lumping scheme should be re-evaluated.

## 4.2 Goodness of fit test statistics

### 4.2.1 Likelihood ratio Goodness of Fit Test

**Full model likelihood** The likelihood of what we shall call the “full model” (observed, unlumped matrix) will be obtained using the estimated probabilities of said matrix, and indicated by  $N_{pr}$ . Any zero-cell elements in this matrix are treated as missing values and thus ultimately do not contribute to the likelihood ratio test statistic. The transition probabilities of the lumped matrix are

The grand sum of the matrix below represents the likelihood of the full model:

$$\begin{aligned}
 F &= N \circ \ln(N_{pr}) \\
 &= \begin{bmatrix} 44 & 46 & 5 & 6 \\ 45 & 43 & 6 & 5 \\ 3 & 2 & 46 & 47 \\ 3 & 6 & 47 & 46 \end{bmatrix} \circ \ln \left( \begin{bmatrix} 0.4356436 & 0.4554455 & 0.049505 & 0.0594059 \\ 0.4545455 & 0.4343434 & 0.0606061 & 0.0505051 \\ 0.0306122 & 0.0204082 & 0.4693878 & 0.4795918 \\ 0.0294118 & 0.0588235 & 0.4607843 & 0.4509804 \end{bmatrix} \right) \\
 &= \begin{bmatrix} -36.56096 & -36.17804 & -15.02841 & -16.94017 \\ -35.48058 & -35.85855 & -16.82016 & -14.92841 \\ -10.45907 & -7.783641 & -34.791 & -34.53653 \\ -10.57908 & -16.99928 & -36.41678 & -36.63125 \end{bmatrix}
 \end{aligned}$$

Then the grand sum of this matrix produces a likelihood for the full model of

$$F_s = -395.9919.$$

**Reduced model likelihood** The likelihood of what we shall call the “reduced model” (lumped matrix) is based upon the transition probabilities in the lumped matrix, indicated by  $M_{pr}$ ; the structure of this matrix is based upon the dimension of the original, unlumped matrix and the concept that each transition has equal probability within each lump. For the actual lumped matrix, these can be obtained using the following expression:

$$M_p = (M \oslash m_i) \oslash S_{pl}$$

Where  $S_{pl}$  is the matrix of states contributing to each lump, repeated for as many rows as there are lumps. It is found by first obtaining the row sums of the lumping matrix,  $K$ , which represents the number of states contributing to each lump,  $K_s$ :

$$K_s = \begin{bmatrix} 2 \\ 2 \end{bmatrix}$$

The transpose of  $K_s$  is then repeated  $m=2$  times, such that  $K_s^T = \begin{bmatrix} 2 & 2 \end{bmatrix}$  such that

$$S_{pl} = \begin{bmatrix} 2 & 2 \\ 2 & 2 \end{bmatrix}$$

Then, considering that each lump is comprised of 2 states, the  $(n \times n)$  matrix of these probabilities,  $M_{pr}$ , is found as

$$M_{pr} = K^T \cdot ((M \oslash m_i) \oslash S_{pl}) \cdot K$$

Once again, any zero-cell elements are treated as missing values and do not contribute to the overall test statistic. Then the likelihood of the reduced model is found in like manner as that for the full model, such that the grand sum of the following matrix represents the likelihood of the reduced model:

$$R = N \circ \ln(M_{pr})$$

Continuing our example, the “equal probabilities per lump” in matrix form is

$$\begin{aligned}
M_p &= (M \oslash m_i) \oslash S_{pl} \\
&= \left( \begin{bmatrix} 178 & 22 \\ 14 & 186 \end{bmatrix} \oslash \begin{bmatrix} 200 \\ 200 \end{bmatrix} \right) \oslash \begin{bmatrix} 2 & 2 \\ 2 & 2 \end{bmatrix} \\
&= \begin{bmatrix} 0.445 & 0.055 \\ 0.035 & 0.465 \end{bmatrix}
\end{aligned}$$

Then the matrix of probabilities is then found as

$$\begin{aligned}
M_{pr} &= K^T \cdot \left( (M \oslash m_i) \oslash S_{pl} \right) \cdot K \\
&= \begin{bmatrix} 1 & 0 \\ 1 & 0 \\ 0 & 1 \\ 0 & 1 \end{bmatrix} \cdot \left( \left( \begin{bmatrix} 178 & 22 \\ 14 & 186 \end{bmatrix} \oslash \begin{bmatrix} 200 \\ 200 \end{bmatrix} \right) \oslash \begin{bmatrix} 2 & 2 \\ 2 & 2 \end{bmatrix} \right) \cdot \begin{bmatrix} 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 \end{bmatrix} \\
&= \begin{bmatrix} 0.445 & 0.445 & 0.055 & 0.055 \\ 0.445 & 0.445 & 0.055 & 0.055 \\ 0.035 & 0.035 & 0.465 & 0.465 \\ 0.035 & 0.035 & 0.465 & 0.465 \end{bmatrix}
\end{aligned}$$

The likelihood the reduced model can then be determined:

$$\begin{aligned}
R &= N \circ \ln(M_{pr}) \\
&= \begin{bmatrix} 44 & 46 & 5 & 6 \\ 45 & 43 & 6 & 5 \\ 3 & 2 & 46 & 47 \\ 3 & 6 & 47 & 46 \end{bmatrix} \circ \ln \left( \begin{bmatrix} 0.445 & 0.445 & 0.055 & 0.055 \\ 0.445 & 0.445 & 0.055 & 0.055 \\ 0.035 & 0.035 & 0.465 & 0.465 \\ 0.035 & 0.035 & 0.465 & 0.465 \end{bmatrix} \right) \\
&= \begin{bmatrix} -35.62596 & -37.24533 & -14.50211 & -17.40253 \\ -36.43564 & -34.81628 & -17.40253 & -14.50211 \\ -10.05722 & -6.704814 & -35.22302 & -35.98874 \\ -10.05722 & -20.11444 & -35.98874 & -35.22302 \end{bmatrix}
\end{aligned}$$

Then the grand sum of this matrix is  $R_s = -397.2897$ .

**Overall likelihood ratio test statistic and degrees of freedom** The overall likelihood ratio test statistic,  $L$ , is found in the conventional manner such that

$$\begin{aligned} X_L^2 &= -2 \cdot (R_s - F_s) \\ &= -2(-397.2897 - -395.9919) \\ &= 2.5956352 \\ &\approx 2.60 \end{aligned}$$

The associated degrees of freedom are found using the number of rows (states) in the observed, unlumped matrix and the number of rows (states) in the lumped matrix. Alternately, these expressions can be thought of in terms of columns, as only square matrices are employed in Markov methodology and mathematics. The degrees of freedom,  $DF_L$ , are found by

$$\begin{aligned} DF_L &= n(n - 1) - m(m - 1) \\ &= 4(4 - 1) - 2(2 - 1) \\ &= 10 \end{aligned}$$

These degrees of freedom are empirically demonstrated via simulation. The associated p-value provides evidence of the lumping scheme being a good fit to the data ( $p = 0.9894076 \approx 0.9894$ ).

#### 4.2.2 Pearson Goodness of Fit Test

The observed lumped frequencies are those expressed in the matrix,  $M$ . The expected lumped frequencies,  $E$ , are found as follows; consistent with previous procedures, we shall again treat any zero-cell elements as missing values.

$$\begin{aligned}
E &= M_{pr} \circ n_k \\
&= \begin{bmatrix} 0.445 & 0.445 & 0.055 & 0.055 \\ 0.445 & 0.445 & 0.055 & 0.055 \\ 0.035 & 0.035 & 0.465 & 0.465 \\ 0.035 & 0.035 & 0.465 & 0.465 \end{bmatrix} \circ \begin{bmatrix} 101 \\ 99 \\ 98 \\ 102 \end{bmatrix} \\
&= \begin{bmatrix} 44.945 & 44.945 & 5.555 & 5.555 \\ 44.055 & 44.055 & 5.445 & 5.445 \\ 3.43 & 3.43 & 45.57 & 45.57 \\ 3.57 & 3.57 & 47.43 & 47.53 \end{bmatrix}
\end{aligned}$$

In order to calculate the matrix of chi-square values, we must define another matrix,  $N_{int}$ , which contains the half-integer correction (0.5) for any zero elements in the observed matrix of transition frequencies,  $N$ . The overall corrected Pearson test statistic,  $X_p^2$ , is obtained by taking the grand sum of the following element-wise matrix manipulations:

$$\begin{aligned}
PS_{cell} &= (N_{int} - E)^2 \oslash E \\
&= \left( \begin{bmatrix} 44 & 46 & 5 & 6 \\ 45 & 43 & 6 & 5 \\ 3 & 2 & 46 & 47 \\ 3 & 6 & 47 & 46 \end{bmatrix} - \begin{bmatrix} 44.945 & 44.945 & 5.555 & 5.555 \\ 44.055 & 44.055 & 5.445 & 5.445 \\ 3.43 & 3.43 & 45.57 & 45.57 \\ 3.57 & 3.57 & 47.43 & 47.53 \end{bmatrix} \right)^2 \\
&\quad \oslash \begin{bmatrix} 44.945 & 44.945 & 5.555 & 5.555 \\ 44.055 & 44.055 & 5.445 & 5.445 \\ 3.43 & 3.43 & 45.57 & 45.57 \\ 3.57 & 3.57 & 47.43 & 47.53 \end{bmatrix} \\
&= \begin{bmatrix} 0.0198693 & 0.0247642 & 0.05545 & 0.0389841 \\ 0.0202707 & 0.0252644 & 0.0565702 & 0.0363682 \\ 0.0539067 & 0.5961808 & 0.0040575 & 0.0448738 \\ 0.0910084 & 1.6540336 & 0.0038984 & 0.0431141 \end{bmatrix}
\end{aligned}$$

The grand sum of this matrix results in  $X_p^2 = 2.7652784$ . The Pearson test-statistic has the same number of degrees of freedom as the LRT, again demonstrated empirically. Consistent with the LRT results, the Pearson test's  $p$ -value provides

evidence of the lumping scheme being a good fit to the data ( $p = 0.986421 \approx 0.9864$ ).

#### 4.2.3 The $W$ -score

Let the cell count,  $C_c$  be a  $J_n$ , or the conventional  $n \times n$  matrix of 1s. Then the number of cells from the observed matrix contributing to each lump is found by

$$C_l = K \cdot C_c \cdot K^T$$

Then the mean number of cells per lump,  $N_{ncell}$ , is found by

$$N_{ncell} = K^T \cdot C_l \cdot K$$

The mean cell transition probability per state (for the unlumped matrix ) is then found as

$$M_{probstate} = N_{pr} \oslash N_{ncell}$$

The mean cell probability for the lumped matrix is found as follows

$$P_{lbar} = K \cdot M_{probstate} \cdot K^T$$

For computational purposes, we shall create an  $n \times n$  matrix of the mean lumped transition probabilities

$$P_{lbar,repeat} = K^T \cdot P_{lbar} \cdot K$$

The difference between the mean unlumped transition probabilities and the mean lumped transition probabilities are then obtained as

$$W_d = N_{pr} - P_{lbar,repeat}$$

After the square of the differences are obtained (as below), the grand sum of this matrix of squared differences yields the  $W$ -statistic.

$$W = (W_d)^2$$

Then, continuing our example, we observe the following. The number of cells from the observed matrix contributing to each lump

$$\begin{aligned} C_l &= K \cdot Cell_{count} \cdot K^T \\ &= \begin{bmatrix} 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 \end{bmatrix} \cdot \begin{bmatrix} 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 \end{bmatrix} \cdot \begin{bmatrix} 1 & 0 \\ 1 & 0 \\ 0 & 1 \\ 0 & 1 \end{bmatrix} \\ &= \begin{bmatrix} 4 & 4 \\ 4 & 4 \end{bmatrix} \end{aligned}$$

The mean number of cells per lump,  $N_{ncell}$ , is

$$\begin{aligned} N_{ncell} &= K^T \cdot C_l \cdot K \\ &= \begin{bmatrix} 1 & 0 \\ 1 & 0 \\ 0 & 1 \\ 0 & 1 \end{bmatrix} \cdot \begin{bmatrix} 4 & 4 \\ 4 & 4 \end{bmatrix} \cdot \begin{bmatrix} 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 \end{bmatrix} \\ &= \begin{bmatrix} 4 & 4 & 4 & 4 \\ 4 & 4 & 4 & 4 \\ 4 & 4 & 4 & 4 \\ 4 & 4 & 4 & 4 \end{bmatrix} \end{aligned}$$

The mean cell transition probability per state (for the unlumped matrix ) is then

$$\begin{aligned} M_{probstate} &= N_{pr} \oslash N_{ncell} \\ &= \begin{bmatrix} 0.4356436 & 0.4554455 & 0.049505 & 0.0594059 \\ 0.4545455 & 0.4343434 & 0.0606061 & 0.0505051 \\ 0.0306122 & 0.0204082 & 0.4693878 & 0.4795918 \\ 0.0294118 & 0.0588235 & 0.4607843 & 0.4509804 \end{bmatrix} \oslash \begin{bmatrix} 4 & 4 & 4 & 4 \\ 4 & 4 & 4 & 4 \\ 4 & 4 & 4 & 4 \\ 4 & 4 & 4 & 4 \end{bmatrix} \\ &= \begin{bmatrix} 0.1089109 & 0.1138614 & 0.0123762 & 0.0148515 \\ 0.1136364 & 0.1085859 & 0.0151515 & 0.0126263 \\ 0.0076531 & 0.005102 & 0.1173469 & 0.119898 \\ 0.0073529 & 0.0147059 & 0.1151961 & 0.1127451 \end{bmatrix} \end{aligned}$$



The mean cell probability for the lumped matrix is then

$$\begin{aligned}
P_{lbar} &= K \cdot M_{probstate} \cdot K^T \\
&= \begin{bmatrix} 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 \end{bmatrix} \cdot \begin{bmatrix} 0.1089109 & 0.1138614 & 0.0123762 & 0.0148515 \\ 0.1136364 & 0.1085859 & 0.0151515 & 0.0126263 \\ 0.0076531 & 0.005102 & 0.1173469 & 0.119898 \\ 0.0073529 & 0.0147059 & 0.1151961 & 0.1127451 \end{bmatrix} \\
&\quad \cdot \begin{bmatrix} 1 & 0 \\ 1 & 0 \\ 0 & 1 \\ 0 & 1 \end{bmatrix} \\
&= \begin{bmatrix} 0.4449945 & 0.0550055 \\ 0.0348139 & 0.4651861 \end{bmatrix}
\end{aligned}$$

The  $n \times n$  matrix of the mean lumped transition probabilities is then

$$\begin{aligned}
P_{lbar,repeat} &= K^T \cdot P_{lbar} \cdot K \\
&= \begin{bmatrix} 1 & 0 \\ 1 & 0 \\ 0 & 1 \\ 0 & 1 \end{bmatrix} \cdot \begin{bmatrix} 0.4449945 & 0.0550055 \\ 0.0348139 & 0.4651861 \end{bmatrix} \cdot \begin{bmatrix} 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 \end{bmatrix} \\
&= \begin{bmatrix} 0.4449945 & 0.4449945 & 0.0550055 & 0.0550055 \\ 0.4449945 & 0.4449945 & 0.0550055 & 0.0550055 \\ 0.0348139 & 0.0348139 & 0.4651861 & 0.4651861 \\ 0.0348139 & 0.0348139 & 0.4651861 & 0.4651861 \end{bmatrix}
\end{aligned}$$

The difference between the mean unlumped transition probabilities and the mean lumped transition probabilities are then

$$\begin{aligned}
W_d &= N_{pr} - P_{lbar,repeat} \\
&= \begin{bmatrix} 0.4356 & 0.4554 & 0.0495 & 0.0594 \\ 0.4545 & 0.4343 & 0.0606 & 0.0505 \\ 0.0306 & 0.0204 & 0.4694 & 0.4796 \\ 0.0294 & 0.0588 & 0.4608 & 0.4510 \end{bmatrix} \\
&\quad - \begin{bmatrix} 0.4449945 & 0.4449945 & 0.0550055 & 0.0550055 \\ 0.4449945 & 0.4449945 & 0.0550055 & 0.0550055 \\ 0.0348139 & 0.0348139 & 0.4651861 & 0.4651861 \\ 0.0348139 & 0.0348139 & 0.4651861 & 0.4651861 \end{bmatrix} \\
&= \begin{bmatrix} -0.009351 & 0.010451 & -0.005501 & 0.0044004 \\ 0.009551 & -0.010651 & 0.0056006 & -0.0045 \\ -0.004202 & -0.014406 & 0.0042017 & 0.0144058 \\ -0.005402 & 0.0240096 & -0.004402 & -0.014206 \end{bmatrix}
\end{aligned}$$

The differences are obtained and the grand sum of this matrix of squared differences yields the  $W$ -score.

$$W = (W_d)^2 = \begin{bmatrix} 0.0000874 & 0.0001092 & 0.0000303 & 0.0000194 \\ 0.0000912 & 0.0001134 & 0.0000314 & 0.0000203 \\ 0.0000177 & 0.0002075 & 0.0000177 & 0.0002075 \\ 0.0000292 & 0.0005765 & 0.0000194 & 0.0002018 \end{bmatrix}$$

The resulting grand sum is  $W_s = 0.0004449$  or  $4.45 \times 10^{-4}$ .

## C. Results

### 1. Summary statistics

In general, the LRT had higher values for mean, median and variance compared to the Pearson, within the same sample size. As the sample size increased, the LRT and Pearson statistics have similar summary statistics. This finding is consistent with known behavior of the LRT and Pearson statistics when evaluating categorical data.

We used Chain 1 to investigate the half-integer correction and distributional properties; only the first two rows have changed from the original actual transition matrix (Figure 1). Note that under Chain 1 the null hypothesis is true, therefore the lumped matrix should provide a better fit to the data compared to the unlumped matrix.

### 2. Distribution of test statistics and degrees of freedom

Using Chain 1, we investigated the distributional properties of the proposed test statistics. For all sample sizes and test statistics, the respective histograms suggest that the statistics follow an asymptotic Chi-square distribution with 10 degrees of freedom (DF). However, this is more apparent with the large sample size of  $N=5,000$ .

Then the DF can be found by

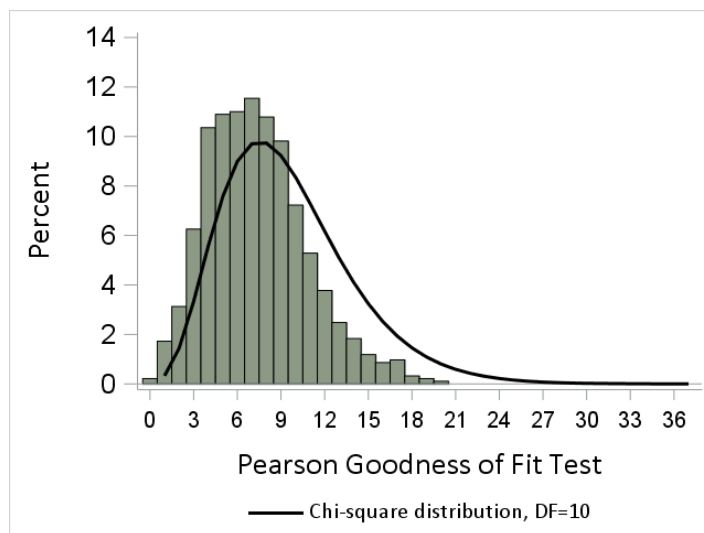
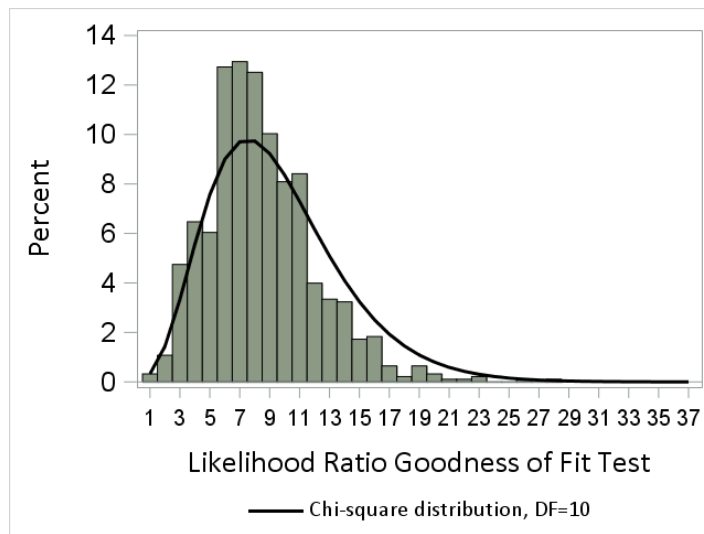
$$\begin{aligned} DF &= ((n \times n) - n) - ((m \times m) - m) \\ &\equiv (n^2 - n) - (m^2 - m) \\ &\equiv n(n - 1) - m(m - 1) \end{aligned}$$

Per our example, we let  $m = 2$  and  $n = 4$ ; then the DF can be found as:

$$\begin{aligned} DF &= n(n - 1) - m(m - 1) \\ &= 4(4 - 1) - 2(2 - 1) \\ &= 4(3) - 2(1) \\ &= 12 - 2 \\ &= 10 \end{aligned}$$

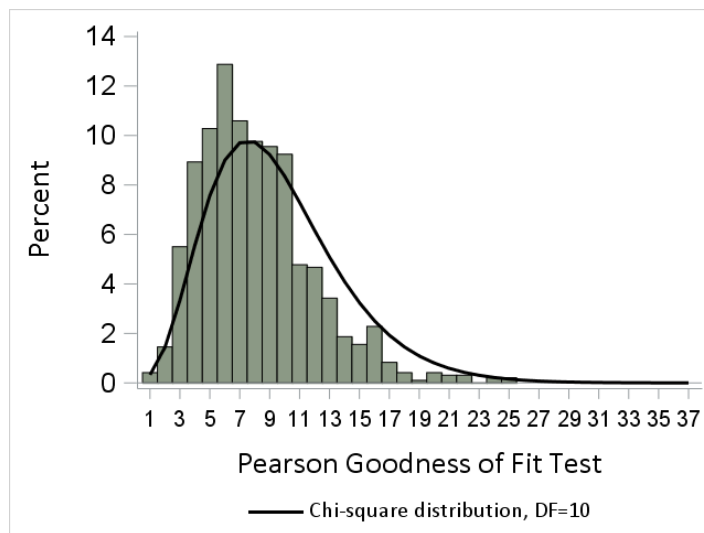
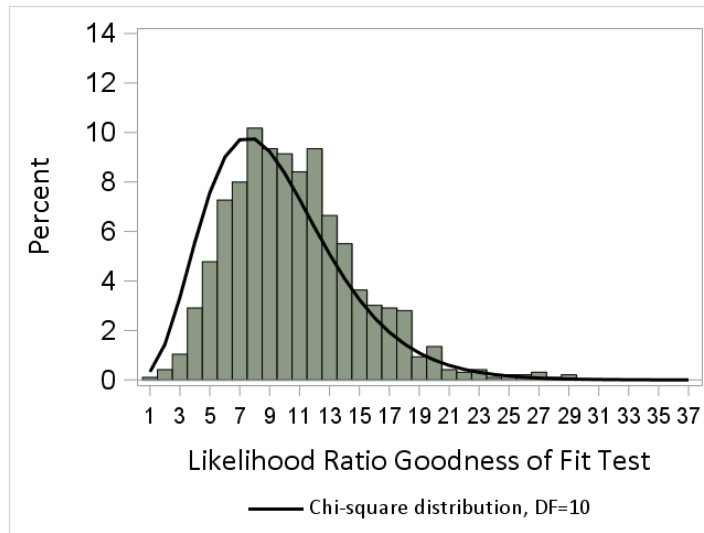
The test statistics' degrees of freedom were depicted using only those simulated matrices (iterations) which were lumpable according to the first scheme (Chain 1); this is because it is only on these matrices that we can appropriately apply the test. While other degrees of freedom were investigated (i.e., 8, 9 and 10), the demonstrated degrees of freedom (degrees of freedom=10) appear to be the most appropriate, based on two key pieces of support. First, the overlay of the curve for a Chi-square distribution with 10 degrees of freedom, and second, the performance of the Kolmogorov-Smirnov and Anderson-Darling tests (Figures 9-12, Table 6). For the two smaller sample sizes (50, 200), the KS test does not support the appropriateness Chi-square distribution with DF=10; however, results for the larger sample size (5000) suggest that this distribution is appropriate to the data and so the distribution of the test statistics are asymptotically Chi-square, with DF=10. It is possible that these results are an artifact of the known limitations of these tests in that they are sensitive to outliers and sample size.

**Figure 9:** Histograms of the Likelihood Ratio and Pearson test statistics; N=50 for 927 iterations



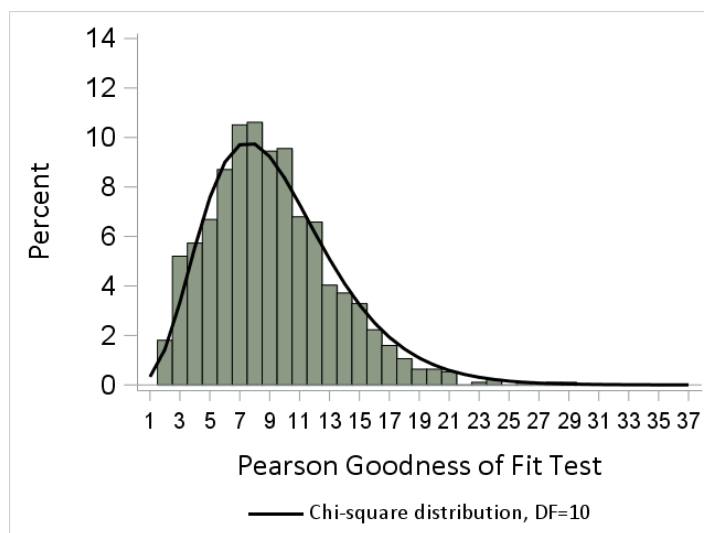
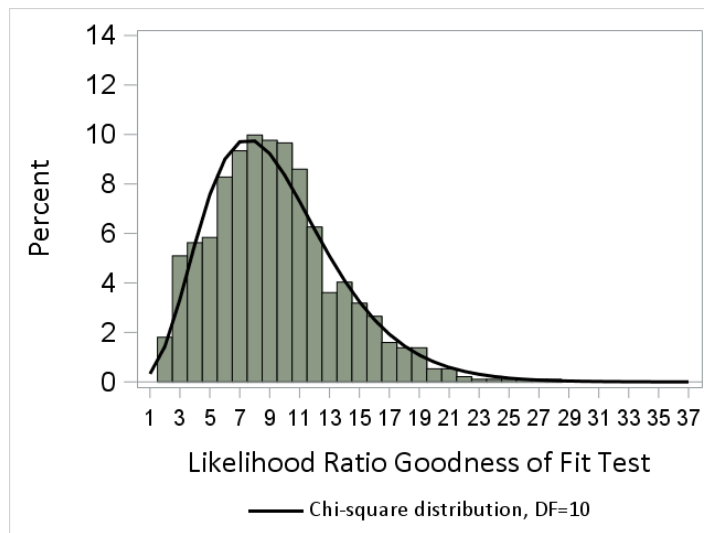
**Figure 10:** Histograms of the Likelihood Ratio and Pearson test statistics; N=200 for 970 iterations

---



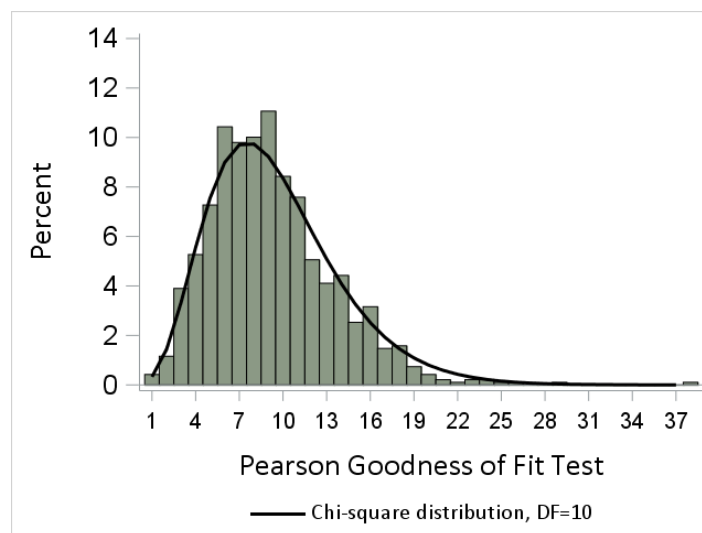
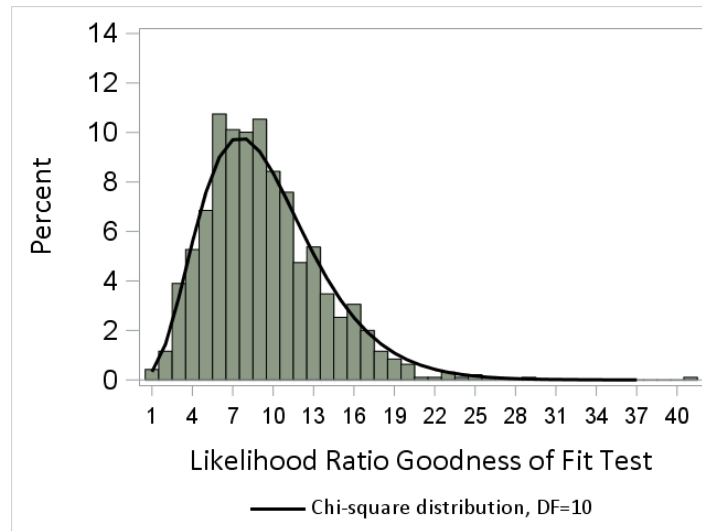
**Figure 11:** Histograms of the Likelihood Ratio and Pearson test statistics; N=1000 for 942 iterations

---



**Figure 12:** Histograms of the Likelihood Ratio and Pearson test statistics; N=5000 for 949 iterations

---





## D. Discussion

### 1. Goodness of Fit

These tests were developed under the notion of performing a two-stage process: first, determining if the chain was lumpable according to a given scheme; and second, if it is lumpable, determining if lumping is the best fit to the data. In that sense, the proposed GOF tests are summarized based only on those lumpable chains, reflecting the appropriateness of performing the GOF.

Both GOF statistics are structured to have the same hypotheses: the null hypothesis is that the lumped matrix (reduced) is a better fit to the data, versus the alternative that the unlumped (observed or saturated) matrix is a better fit to the data.

Therefore, a large  $p$ -value is desirable in order to demonstrate that the lumped (reduced) matrix is the better fitting lumping scheme. The summary statistics and comparison figures (presented and discussed in Section C.3) indicate the similar performance of the two GOF statistics. As anticipated, values become more consistent between the statistics with an increase in sample size.

### 2. $W$ -score

The  $W$ -score describes a property of the true transition matrix. As mentioned earlier, it was developed to assist in the examination of different chains for the investigation of the proposed test statistics. Its purpose was to demonstrate that the starting 4 chains had varying levels of lumpability (Tables 5, 6). It was

summarized for all matrices simulated from each of the 4 chains, as it is intended for anticipating the results of the Test of Lumpability (Stage 1). Importantly, it was difficult to predict the value of  $W$ , itself, because moving a zero-cell, very small probability, or even changing a probability slightly changed  $W$  in unpredictable ways.

**Table 5:** Probabilities used for simulation

Chain	"True" transition probabilities, $B_0$	$W$ - score	Notes
1*	$\begin{bmatrix} 0.48 & 0.48 & 0.02 & 0.02 \\ 0.48 & 0.48 & 0.02 & 0.02 \\ 0.03125 & 0.03125 & 0.46875 & 0.46875 \\ 0.03125 & 0.03125 & 0.46875 & 0.46875 \end{bmatrix}$	0	Designed to be perfectly lumpable
2	$\begin{bmatrix} 0.44 & 0.46 & 0.04 & 0.06 \\ 0.4475 & 0.4275 & 0.0625 & 0.0625 \\ 0.03 & 0.02 & 0.47 & 0.48 \\ 0.03125 & 0.03125 & 0.46875 & 0.46875 \end{bmatrix}$	$2.7 \times 10^{-4}$	Rows 1-3 had changes in entries; row 4 remained the same
3	$\begin{bmatrix} 0.44 & 0.46 & 0.04 & 0.06 \\ 0.4475 & 0.4275 & 0.0625 & 0.0625 \\ 0.029 & 0.021 & 0.46 & 0.49 \\ 0.027 & 0.0555 & 0.47 & 0.4475 \end{bmatrix}$	$6.4 \times 10^{-4}$	
4	$\begin{bmatrix} 0.48 & 0.48 & 0.02 & 0.02 \\ 0.4375 & 0.4375 & 0.0625 & 0.0625 \\ 0.03125 & 0.03125 & 0.46875 & 0.46875 \\ 0.03125 & 0.03125 & 0.46875 & 0.46875 \end{bmatrix}$	$9.0 \times 10^{-4}$	Entries in Row 2 are different from that in chain 1

**Table 6:** Summary of proposed test statistics for Chain 1 lumpable matrices, lumpable matrices only

Simulation sample size	Likelihood Ratio Test			
	50	200	1000	5000
<i>No. of iterations (lumpable matrices), N (%)</i>	927 (92.7)	970 (97.0)	942 (94.2)	949 (94.9)
<i>No. of matrices with <math>\geq 1</math> zero-cell, N (%)</i>	92.7 (100)	904 (93.2)	17 (1.80)	0 (0)
<i>No. of zero-cells<sub>‡</sub>, Mean (SD)</i>	5.76 (1.23)	2.19 (1.22)	1.12 (1.74)	0 (0)
Summary statistics				
<i>Mean of test statistic</i>	8.95	10.84	9.83	9.74
<i>Variance of test statistic</i>	13.16	17.27	18.12	18.36
<i>Median of test statistic</i>	8.46	10.22	9.42	9.20
<i>Minimum of test statistic</i>	1.14	0.45	2.61	1.65
<i>Maximum of test statistic</i>	28.02	31.41	21.61	41.68
Goodness of fit to Chi-square distribution				
<i>DF=10; Kolmogorov-Smirnov, p</i>	<0.0001	<0.0001	0.205	0.083
<i>DF=10; Anderson-Darling, p</i>	<0.0001	<0.0001	0.250	0.089
<i>Choice of lumped matrix (<math>p&gt;0.05</math>), DF=10; N (%)</i>	912 (98.38)	912 (94.02)	903 (95.86)	919 (96.84)
<i>Choice of Unlumped matrix (<math>p&lt;0.05</math>), DF=10; N (%)</i>	15 (1.62)	58 (5.98)	39 (4.14)	30 (3.16)
Simulation sample size	Pearson Test			
	50	200	1000	5000
<i>No. of iterations (lumpable matrices), N (%)</i>	927 (92.7)	970 (97.0)	942 (94.2)	949 (94.9)
<i>No. of matrices with <math>\geq 1</math> zero-cell, N (%)</i>	92.7 (100)	904 (93.2)	17 (1.80)	0 (0)
<i>No. of zero-cells<sub>‡</sub>, Mean (SD)</i>	5.76 (1.23)	2.19 (1.22)	1.12 (1.74)	0 (0)
Summary statistics				
<i>Mean of test statistic</i>	7.87	8.39	9.59	9.70
<i>Variance of test statistic</i>	11.71	12.90	17.24	17.90
<i>Median of test statistic</i>	7.56	7.74	9.12	9.18
<i>Minimum of test statistic</i>	0.71	0.45	2.14	1.64
<i>Maximum of test statistic</i>	20.30	26.86	29.75	38.18
Goodness of fit to Chi-square distribution				
<i>DF=10; Kolmogorov-Smirnov, p</i>	<0.0001	<0.0001	0.049	0.055
<i>DF=10; Anderson-Darling, p</i>	<0.0001	<0.0001	0.018	0.055
<i>Choice of lumped matrix (<math>p&gt;0.05</math>), DF=10; N (%)</i>	922 (99.46)	961 (99.07)	909 (96.5)	922 (97.1)
<i>Choice of Unlumped matrix (<math>p&lt;0.05</math>), DF=10; N (%)</i>	5 (0.54)	9 (0.93)	33 (3.50)	27 (2.85)

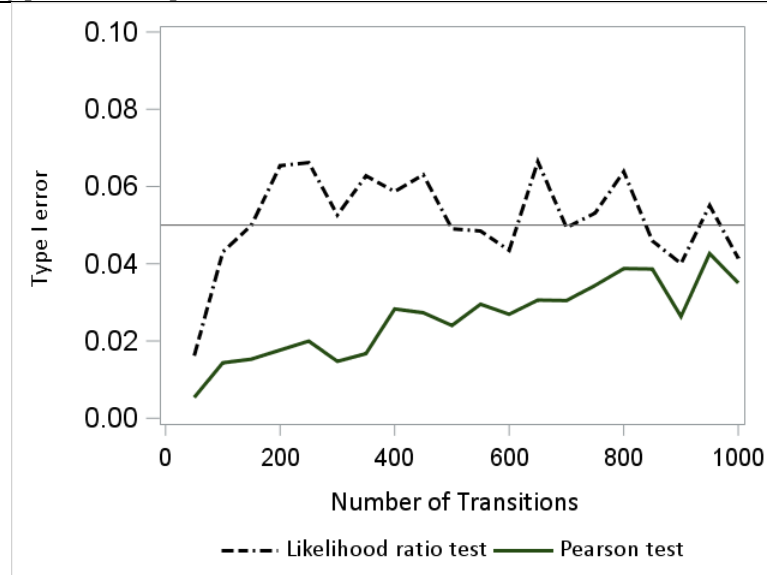
‡In the unlumped, observed matrix  
‡Measure of difference between observed and lumped matrices

### 3. Performance of proposed goodness of fit tests

#### 3.1 Type I error

Type I error (chain 1) was investigated for our proposed tests for each sample size. The LRT has greater type I error than the Pearson test. However, the differences are observed to be greater at smaller sample sizes, with results becoming more consistent with each other as the sample size increases (Figure 13). For the LRT, the type I error hovers about the desired 0.05 level most of the time. The Pearson test is more conservative, with type I error consistently smaller than 0.05 for all sample sizes, although it approaches that 0.05-level as the sample size increases.

**Figure 13:** Type 1 error plot, Chain 1

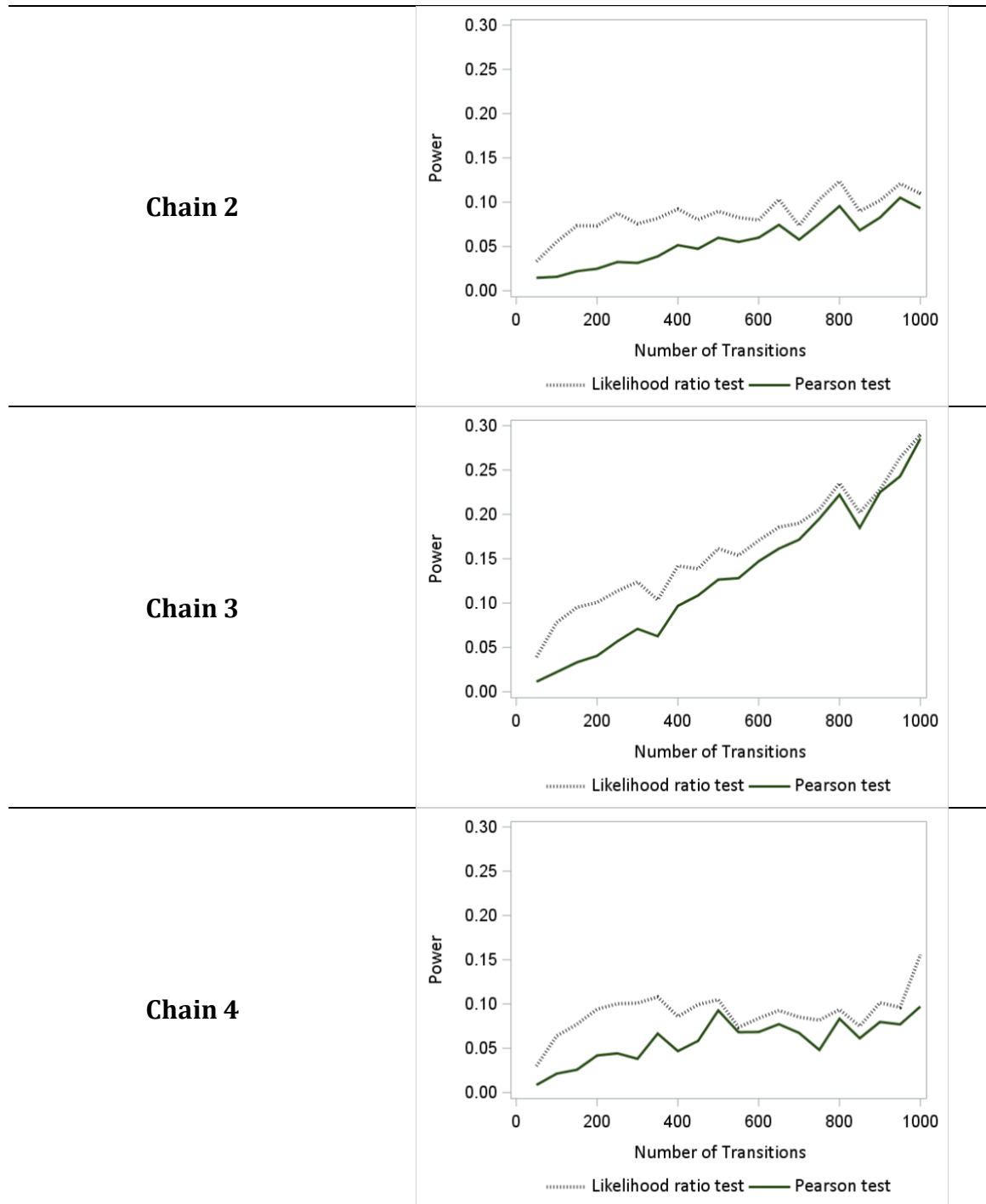


#### 3.2 Power

Power (chains 2-4) was investigated for our proposed tests, with each chain and for each sample size. For all chains, the LRT has more power than the Pearson tests

(Figure 14). Again, the differences are observed at smaller sample sizes, with results becoming more consistent with each other as the sample size increases.

**Figure 14:** Power plots, chains 2-4



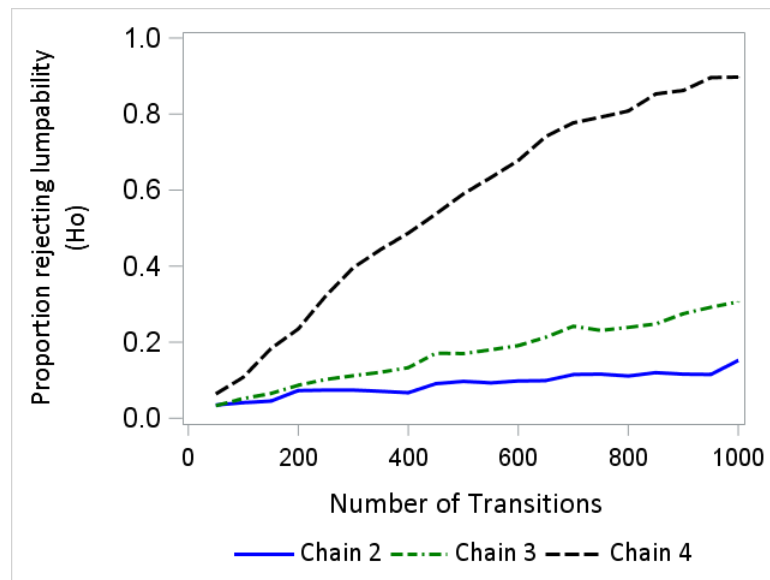
### 3.3 Comparing the Chains

**Stage 1** Larger values of  $W$  indicate we are more likely to reject lumpability.

Therefore, the further we are from a perfectly lumpable matrix ( $W=0$ ), we are more likely to reject the Test of Lumpability for that chain and scheme combination.

These patterns were observed in our simulated data, again noting that the chains are ordered by increasing value of  $W$ -score (Figure 15). As  $W$  increases, so does the tendency to reject the Test of Lumpability, and this occurs with increasing sample size. When comparing the chains, themselves, we found that they do differ by  $W$  scores, and this is particularly observed as the sample size increases for each chain. Chain 4, for instance, as the highest  $W$  score and its proportion of rejecting the test of lumpability increases with sample size at a greater rate than Chains 2 and 3; in fact, the proportion of rejected tests approaches 1.0 as sample size increases.

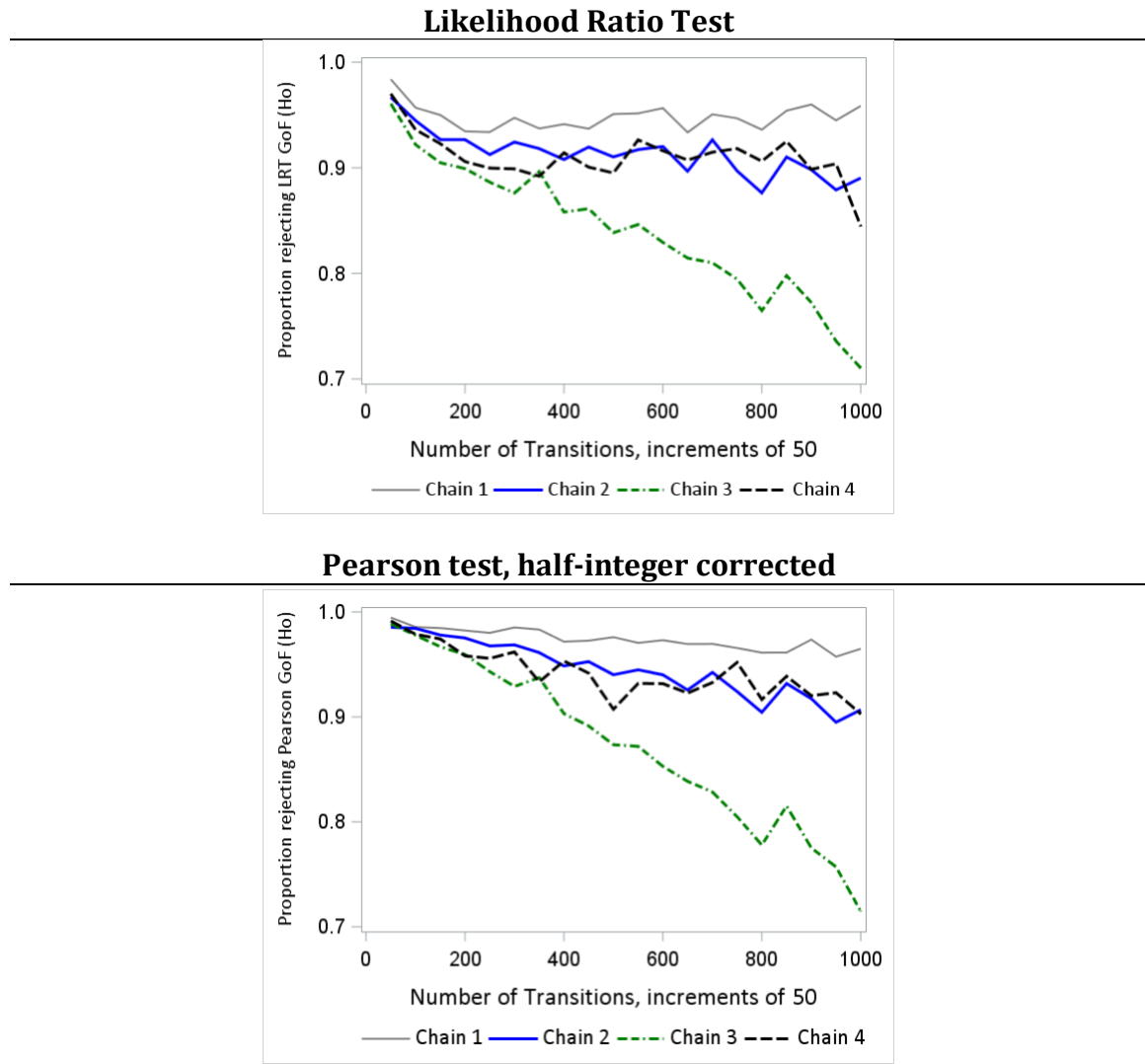
**Figure 15:** Stage 1, rejecting the test of lumpability\*



\*1000 iterations per sample size

**Stage 2** For both tests and all chains, as sample size increases, so does the probability of choosing the unlumped chain. It is interesting that the chains do not perform consistently relative to the  $W$ -scores. That is, for both tests, Chain 1 is most consistent within the chain, across sample sizes, and whose proportion of rejecting the null hypothesis hovers closes to 1.0 as sample size increases. Chain 2 has the next smallest  $W$ - score, and is closest to Chain 1's curve; however, the next closest line in the figure is from Chain 4, which has the highest  $W$ -score (Figure 16).

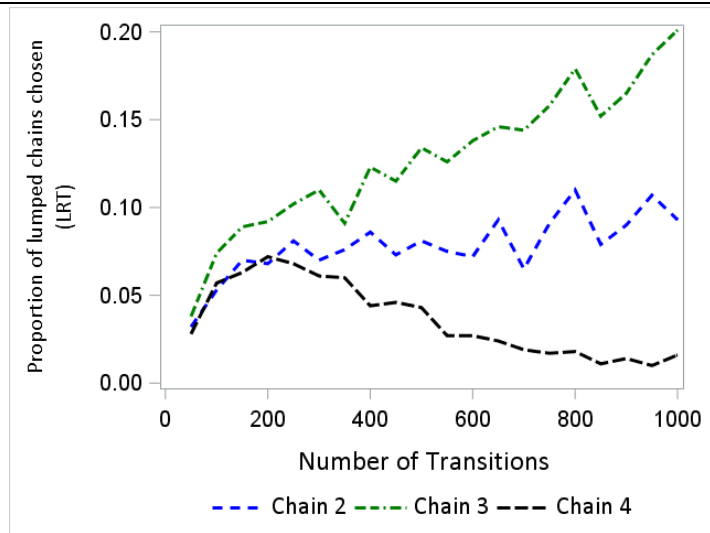
**Figure 16:** Stage 2, Choosing unlumped after determining lumpability



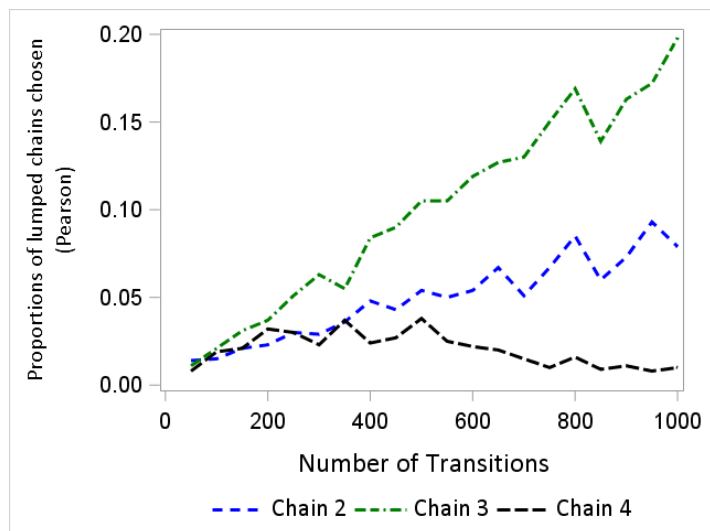
**Choosing a lumped matrix after both testing stages** Choosing the lumped chain varies with the chain, itself. Chain 1 fluctuates around the 5% rate of choosing the lumped chain as the sample size increases (Figure 17). With increasing sample size, Chains 2 and 3 exhibit a pattern of choosing the lumped chain after both stages of testing and under both GOF tests. Likewise, Chain 4 demonstrates a pattern of not choosing the lumped chain with increasing sample size. Because the chains are ordered by increasing  $W$ -scores, it may seem counterintuitive that we do not observe the same pattern in these figures (in terms of rejection) as we do for Figure 16. However, this is due to the fact that we have a 2-stage process and the non-lumpable chains are not considered here because they are filtered out in Stage 1.



**Figure 17: Two-stage process, choosing lumped after both tests**  
**Likelihood Ratio Test**



**Pearson test, half-integer corrected**



## E. Summary, Discussion and Conclusion

**Summary** This chapter described how we extended current lumpability methodology by developing a GOF test with two formulations. In doing so, we estimated transition probabilities and demonstrated that our proposed test statistics appear to follow a Chi-square distribution whose degrees of freedom depend on the number of states in the original, unlumped chain and the number of lumps in the lumped chain. We also developed a comparator statistic, the  $W$ -score, which describes the lumpability of a matrix relative to one that is perfectly lumpable.

**Discussion** Understanding the nuanced difference between the statistical appropriateness of a test and the best fit to the data at hand drove our investigation of goodness of fit of lumpability. Simulation methods were implemented to assess our proposed statistics' distributions, degrees of freedom and performance.

While most chains were lumpable (Stage 1 testing), there was a difference across the  $W$ -score. Increasing  $W$ -scores appear more likely to fail the Test of Lumpability with increasing sample size.

Stage 2 testing resulted in inconsistent test rejection rates, therefore making it difficult to draw a conclusion regarding how the chains perform under each test condition, by the  $W$ -score. This suggests a limitation in the  $W$ -score to the GOF tests and the fact that as a single number, it can only account for so much variation

between the lumped and unlumped matrices. When considering the number of cells that the  $W$ -score is describing, this is not an unreasonable suggestion.

The final pair of figures (Figure 17) addresses our initial question of the worth of statistically-supported lumping. Overall, lumped chains were *not* chosen as frequently in the face of increasing sample size. The fact that there is a decreasing change in choosing a lumped matrix in some cases as the sample size increases suggests that the utility of lumping lies with smaller sample sizes. Therefore, larger sample sizes probably have large enough values in each cell of the transition matrix, and so lumping states together causes loss of information.

All figures comparing test performance between chains demonstrate that the LRT and Pearson tests perform similarly. However, implementation of these GOF tests on actual study data will provide greater insight into their performance and limitations.

Upon establishing the lumpability goodness of fit test, the natural extension is to explore its performance in real data. In order to observe how the data perform a larger sample size and over a longer period of time, we shall apply the Markov methodology to the NARCOMS data. Because the EDSS is the gold standard in MS clinical trials, we will then consider Markov methodology in the CombiRx data. Of interest will be to consider if the Markov models utilize similar lumping schemes and/or utilize similar predictors to produce the Markov chains. We shall also be in the position of evaluating how the methodology performs under differing data patterns, over differing periods of time.

#### ***IV. Applications to NARCOMS Registry Data: PDDS and Lumpability***

##### **A. Introduction**

The overarching goal of this chapter was to investigate disability progression using Markovian methodology to model the PDDS and employing covariates which contribute to the movement between PDDS scores, utilizing data from the North American Research Committee on Multiple Sclerosis (NARCOMS). Our specific goals were to (1) identify and implement useful lumping schemes for PDDS; (2) demonstrate the applicability of the Test of Lumpability for the PDDS; and (3) implement the 2-stage lumpability assessment process. Investigation of the PDDS via Markov methodology was novel, as was the implementation of the Test of Lumpability.

##### **B. Methods**

###### **1. Study Design**

*Study Design* NARCOMS is associated with the Consortium of Multiple Sclerosis Centers (CMSC); it is an active, longitudinal registry for MS participants who provide self-reported health information. Participants submit enrollment information, followed by bi-annual update surveys (by mail or online). Updates provide socio-

demographic and health information. Disability status is recorded using Patient Determined Disease Steps (PDDS) (Hohol et al., 1999), a validated tool used as a surrogate for with the Expanded Disability Status Scale (EDSS) (Learmonth et al., 2013). The PDDS measures disability level from the perspective of the participant (self-report); it is on an ordinal scale from 0 (normal, no disability) to 8 (bedridden) (Marrie & Goldman, 2007). Patients were dichotomized into disability levels according to their PDDS at enrollment. Patients with PDDS  $\leq 4$  were considered Mildly Impaired (MI) (N=1545); patients with PDDS  $\geq 5$  were considered Highly Impaired (HI) (N=502). This stratification was consistent with a recent study on NARCOMS data (Liu et al., 2016).

*Protocol approval and patient consent* All data were de-identified and participants gave informed consent. This study was approved by the institutional review board at the University of Alabama at Birmingham, Birmingham, Alabama.

This was a longitudinal observational study. Primary variables considered from the enrollment survey included date of birth (DOB), gender (male, female), year of MS diagnosis, age at enrollment, age at diagnosis, disease duration at enrollment, PDDS score at enrollment, and PDDS at follow-up (Tables 7 and 8). Follow-up data included in the analyses was PDDS only.

*Participants* Participants included adult male and female MS participants registered with NARCOMS. There were 7,587 participants with enrollment data and at least one follow-up response during the 2007-2012 observation period. Of those, 2,047

(27.0%) had complete PDDS response data; the remainder were excluded (see Inclusion Criteria).

*Inclusion criteria* Participants with baseline enrollment PDDS and biannual PDDS updates between 2007-2012 were included. Only participants with PDDS responses for each survey time during the observation period were considered.

*Exclusion criteria* Participants with any missing responses for PDDS (enrollment and survey updates) were excluded from analysis.

## 2. Preliminary Analysis

Data were summarized using means, standard deviations (SD), frequencies (N) and medians. Group comparisons were performed using Likelihood Ratio (LR) Chi-squared, Analysis of Variance (ANOVA) or Wilcoxon test, and Fisher's exact test, Binomial and Chi-square tests of proportion, as appropriate (Tables 7 and 8).

## 3. Primary Analysis

The primary outcome was the transition of PDDS scores, which was considered at all survey collection points between 2007 and 2012, and at enrollment. Three samples were considered: overall and stratified by level of impairment at enrollment (Mildly, where  $PDDS \leq 4$  ; or Highly, where  $PDDS \geq 5$ ); the associated frequency transition matrices were calculated. Three lumping schemes were

considered to reduce the number of states from the original 9-state chain (Figure 18).

Scheme 1 reflects the original scale for the Disease Steps (DS), the basis of the EDSS; only scores 7 and 8 are collapsed (Hohol et al., 1995; Rizzo et al., 2004). While noting there is no one-to-one correspondence between EDSS and PDDS, Scheme 2 was designed to reflect their similarity and approximate correspondence, where appropriate (Marrie et al., 2006; Marrie & Goldman, 2007) (Figure 18). Scores 1 and 2 were combined because they represent no disability, but mild symptoms; there is not clear correspondence with EDSS scores. A score of 3 indicates some gait disability, but no requirement of assistive devices; this corresponds with an EDSS of 4 and 4.5. PDDS scores of 4, 5, and 6 indicate assistive devices are required for mobility; this corresponds with EDSS scores of 6 and 6.5. PDDS score of 7 corresponds with EDSS scores 7 and 7.5, and indicate a patient is wheelchair-bound. A PDDS score of 8 corresponds with an EDSS of 8 and indicates a patient is bed-bound.

Each scheme was applied to form a new, lumped chain and the Test of Lumpability was performed to compare the original, unlumped chain to the lumped chain. Large  $p$ -values indicate that the chain is lumpable under the given scheme, so that the Markov dependency is retained in the new, smaller chain; smaller  $p$ -values ( $p < 0.05$ ) indicate a chain is not lumpable under that scheme (Jernigan & Baran, 2003). The proposed Chi-square goodness of fit tests comparing the lumped and unlumped chains were performed on the schemes with  $p \geq 0.05$  for the test of lumpability.

#### 4. Secondary Analysis

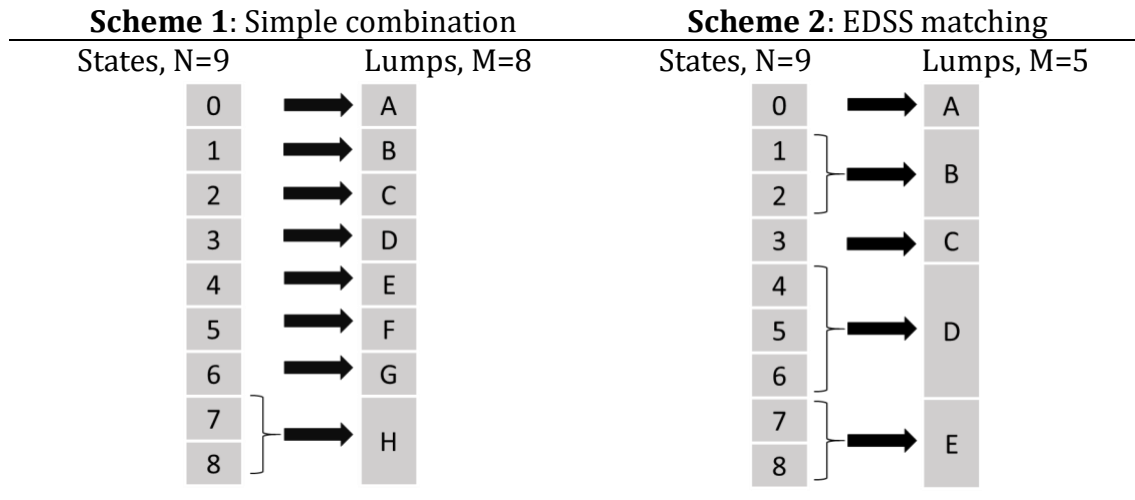
Transition probabilities were estimated for the overall sample and stratified by enrollment group, adjusting for covariates using a proportional odds model.

Random effects (intercept) were included in the model, to account for the heterogeneous nature of the disease between participants. The Newton-Raphson method was used for optimization; Compound Symmetry was used as the covariance structure. Predictors included gender, race, relapse history, age of diagnosis, disease duration at enrollment, enrollment PDDS, and current PDDS to predict next PDDS. For numeric stability and model convergence, Current PDDS and Enrollment PDDS were treated as numeric. When using these lumped “current” PDDS scores, the mean PDDS score was used for prediction. Specifically, for Scheme 1 (Simple combination), all numeric values correspond one-to-one with the original scores, except for Scores 7 and 8, which are combined; in that case, the numeric value was 7.5. For scheme 2 (EDSS Matching), the value of 0 was used for Lump A, 1.5 was used for Lump B, 3 for Lump C, 5 for Lump D and 7.5 for Lump E.

Statistical analyses were performed in SAS V9.4, SAS/IML v14.3 and JMP Pro V14.0 (SAS Institute, Inc., Cary, NC) using an  $\alpha=0.05$  significance level.



**Figure 18:** Lumping schemes for the PDDS



## C. Results

### 1. Preliminary Analyses

**Gender** There were more females than males in both mobility categories, and with each enrollment disability group ( $\chi^2(df=1)=24.1, p<0.0001$ ), following the general pattern of having more females (N=1,549) than males (N=498) overall ( $\chi^2(df=1)=539.6, p<0.0001$ ; Table 7).

**Race** There were more White patients than any other race ( $\chi^2(df=2)=3499.3, p<0.0001$ ); the number of patients who are African American or responded “other” are not statistically different. These proportions were consistent between the disability groups (MI:  $\chi^2(df=2)=2671.8, p<0.0001$ ; HI:  $\chi^2(df=2)=827.9, p<0.0001$ ). There was no association between race and enrollment disability group ( $p=0.2767$ ; Table 7).

**Marital status** At enrollment, more patients were Married or Cohabiting, than Currently Single or Never Married ( $\chi^2(df=2)=1489.2, p<0.0001$ ); the number of patients who are Currently Single or Never Married are not statistically different from each other. This pattern remains true for each disability group (MI:  $\chi^2(df=2)=1091.2, p<0.0001$ ; HI:  $\chi^2(df=2)=400.2, p<0.0001$ ). There was no association between enrollment marital status and disability group ( $p=0.1052$ ; Table 7).

**Age** On average, patients with MI were younger than HI patients ( $p<0.0001$ ) although the age of diagnosis is not statistically different between the disability groups ( $p=0.1221$ ). At enrollment, patients in the HI group have approximately 5-years' longer disease duration, on average, compared to those patients with MI ( $p<0.0001$ ). At the start of the observation period, 2007, patients with MI have shorter disease duration compared those with HI ( $F(1, 2045)=145.8, p<0.0001$ ; Tables 7 and 8).

**Relapse history** More participants have reported experiencing a previous relapse (versus not experiencing relapse) for both disability groups, although a small percentage either did not respond to this question or were unsure. There is some evidence of an association between relapse experience and enrolment disability group ( $\chi^2(df=2)=14.4, p=0.0007$ ). Overall, more patients experienced relapse at enrollment than otherwise ( $\chi^2(df=2)=2326.6, p<0.0001$ ; Table 8).

**Disability** There are more patients in the MI disability group than the HI group, indicating that more patients enroll with higher disability than lower disability,

according to this grouping strategy ( $\chi^2(N=502, df=1)=531.4, p<0.001$ ). Within each disability group, the proportions between the PDDS scores vary. For the MI group, PDDS=2 appears with the lowest frequency; the remaining scores (0, 1, 3, 4) are similar in proportion ( $\chi^2(N=1545, df=4)=113.0, p<0.0001$ ). For the HI group, the proportion of patients in each score decreases as the score increases ( $\chi^2(df=2)=74.9, p<0.0001$ ; Table 8).

**Table 7:** Summary of cohort demographic characteristics at enrollment (N=2,047)

Demographics		Disability level at enrollment			p-value*
		Overall 2,047	Mildly Impaired† 1545 (75.5)	Highly Impaired‡ 502 (24.5)	
<b>Gender, N (%)</b>	Female	1549 (75.7)	1211 (78.4)	338 (67.3)	<0.0001 <sub>1</sub>
	Male	498 (24.3)	334 (21.6)	164 (32.7)	
<b>Race, N (%)</b> (23 missing: 18 for MI, 5 for HI)	African American	27 (1.3)	17 (1.1)	10 (2.0)	0.2767 <sub>2</sub>
	White	1929 (95.3)	1491 (95.7)	468 (94.2)	
	Other	68 (3.4)	49 (3.2)	19 (3.8)	
<b>Marital Status, N (%)</b>	Currently single <sup>A</sup>	293 (14.3)	217 (14.1)	76 (15.1)	0.1052 <sub>3</sub>
	Married/cohabitating	1505 (73.5)	1127 (72.9)	378 (75.3)	
	Never married <sup>B</sup>	249 (12.2)	201 (13.0)	48 (9.6)	
<b>Age (years)</b>	Mean(SD) <sup>C</sup>	47.6 (9.2)	46.4 (9.0)	51.3 (8.8)	<0.0001 <sub>4</sub>
	Median (Min, Max)	48.0 (20.0, 76.0)	47.0 (20.0, 76.0)	52.0 (22.0, 75.0)	
<b>Age at diagnosis (years)</b>	Mean(SD) <sup>C</sup>	37.5 (9.4)	37.6 (9.3)	37.3 (9.7)	0.5231 <sub>5</sub>
	Median (Min, Max)	38.0 (13.0, 65.0)	38.0 (13.0, 64.0)	38.0 (15.0, 65.0)	

†Mildly impaired if PDDS≤4 ‡Highly impaired if PDDS≥5

\*Tests performed on between disability levels on means and proportions

<sup>A</sup>Currently single divorced, widowed, separated) <sup>B</sup>Never married, currently single <sup>C</sup>Standard deviation<sub>1</sub>LR  $\chi^2$  (1, 2047)=24.1 <sub>2</sub>LR  $\chi^2$  (2, 2024)=2.6 <sub>3</sub>LR  $\chi^2$  (2, 2047)=4.5 <sub>4</sub>F(1, 2045)=113.9 <sub>5</sub>F(1, 2045)=0.4

**Table 8:** Summary of cohort clinical characteristics at enrollment (N=2,047)

Clinical characteristics		Disability level at enrollment			p-value*
		Overall 2,047	Mildly Impaired† 1545 (75.5)	Highly Impaired‡ 502 (24.5)	
<b>Disease Duration</b> (years)	Mean(SD) <sub>1</sub>	9.1 (8.3)	7.8 (7.7)	13.0 (8.9)	<0.0001 <sub>2</sub>
	Median (Min, Max)	7.0 (0, 50.0)	5.0 (0, 50.0)	12.0 (0, 47.0)	NA
<b>Disease Duration in 2007</b> (years)	Mean(SD) <sub>1</sub>	15.8 (8.9)	14.5 (8.3)	19.8 (9.5)	<0.0001 <sub>3</sub>
	Median (Min, Max)	14.0 (1.0, 60.0)	12.0 (1.0, 60.0)	19.0 (1.0, 54.0)	NA
<b>Relapse history</b> , N (%) (14 missing, 7 per disability level)	Yes	1701 (83.7)	1314 (85.4)	387 (78.2)	0.0007 <sub>4</sub>
	No	220 (10.8)	145 (9.4)	75 (15.1)	
	Unsure	112 (5.5)	79 (5.1)	33 (6.7)	
<b>PDDS (Numeric)</b>	Mean(SD) <sub>1</sub>	2.9 (2.1)	2.0 (1.5)	5.7 (0.8)	NA
	Median (Min, Max)	3.0 (0, 7.0)	2.0 (0, 4.0)	5.0 (5.0, 7.0)	
	0=Normal	395 (19.3)	395 (25.6)	NA	
	1=Mild disability	291 (14.2)	291 (18.8)	NA	
	2=Moderate disability	156 (7.6)	156 (10.1)	NA	
<b>PDDS, N (%)</b>	3=Gait disability	361 (17.6)	361 (23.4)	NA	NA
	4=Early cane	342 (16.7)	342 (22.1)	NA	
	5=Late cane	257 (12.5)	NA	257 (51.2)	
	6=Bilateral support	138 (6.7)	NA	138 (27.5)	
	7=Wheelchair	107 (5.2)	NA	107 (21.3)	
	8=Bedridden	0 (0)	NA	0 (0)	

†Mildly impaired if PDDS≤4 ‡Highly impaired if PDDS≥5

\*Tests performed on between disability levels on means and proportions

<sub>1</sub>Standard deviation<sub>2</sub>F(1, 2045)=161.2 <sub>3</sub>F(1, 2045)=145.8 <sub>4</sub>LR  $\chi^2$  (2, 2047)=14.4

## 2. Primary Analyses

Without accounting for PDDS score at enrollment or other covariates, the frequency transition matrix depicts relative consistency of the PDDS score over time; this is true whether the chain is lumped or unlumped in the overall sample. That is, most values are concentrated about the diagonal of the matrix (Tables 9, 10, 11). The 2,047 patients account for a total of 22,517 transitions across 10 time points. The chain for the overall sample was not lumpable under either lumping scheme (all  $p < 0.05$ ; Table 18). Examination of the state diagram of the overall depicts transitions between PDDS scores and provides a visual description of the complex, non-linear nature of the outcome (Figure 19).

**Mildly Impaired at Enrollment** In this group, 1,545 patients account for a total of 16,995 transitions across 10 time points (Table 12). Similar to the transition matrix for the overall sample, transitions are concentrated about the diagonal. The chain was not lumpable under either of the lumping schemes (all  $p < 0.05$ ; Tables 12-14, 18). Therefore, neither GOF test was performed for this subset.

**Highly Impaired at Enrollment** In this group, 502 patients account for a total of 5,522 transitions across 10 time points (Table 14). Consistent with the overall sample and the MI transition matrices, transitions are concentrated about the diagonal. The unlumped chain for the HI group was lumpable according to Scheme 1 only; it was not lumpable according to the other proposed scheme (Tables 14-17; 18; Figure 20).

The unlumped chain for the HI subset was lumpable only according to Scheme 1 (Simple Combination) (Figure 20). According to the Chi-square goodness of fit test, this lumping scheme produced a chain that was not a better fit to the data, compared to the unlumped chain. The conclusions between the LRT ( $\chi^2(df = 16) = 3295.44, p < 0.0001$ ) and Pearson formulations ( $\chi^2(df = 16) = 3925.85, p < 0.0001$ ) of the test are consistent.

**Table 9** : Unadjusted, overall transition matrices (N=2,047)

<i>Frequencies</i>										
	0	1	2	3	4	5	6	7	8	
0	2362	567	72	32	7	4	2	1	0	3047
1	529	1963	353	302	41	7	5	2	0	3202
2	58	354	693	291	65	6	6	1	0	1474
3	27	276	283	1478	349	26	12	1	0	2452
4	7	37	56	269	2506	515	52	8	2	3452
5	4	1	10	19	392	1973	352	26	1	2778
6	2	4	5	13	32	238	2175	284	6	2759
7	1	2	2	2	3	18	192	2882	49	3151
8	0	0	0	0	2	1	3	37	159	202
<i>Probabilities</i>										
	0	1	2	3	4	5	6	7	8	
0	0.782	0.19	0.02	0.01	0.002	0.001	0.001	0.0003	0	
1	0.173	0.61	0.11	0.09	0.01	0.002	0.002	0.001	0	
2	0.044	0.24	0.47	0.20	0.04	0.004	0.004	0.001	0	
3	0.015	0.11	0.12	0.60	0.14	0.01	0.005	0.0004	0	
4	0.0026	0.01	0.02	0.08	0.73	0.15	0.02	0.002	0.001	
5	0.0017	0.0004	0.004	0.01	0.14	0.71	0.13	0.01	0.0004	
6	0.0018	0.001	0.002	0.005	0.01	0.09	0.79	0.10	0.002	
7	0.0003	0.001	0.001	0.001	0.001	0.01	0.06	0.91	0.02	
8	0	0	0	0	0.01	0.01	0.01	0.18	0.79	

**Table 10:** Scheme 1: Unadjusted lumped transition matrix, overall (N=2,047; 9 states to 8 lumps)

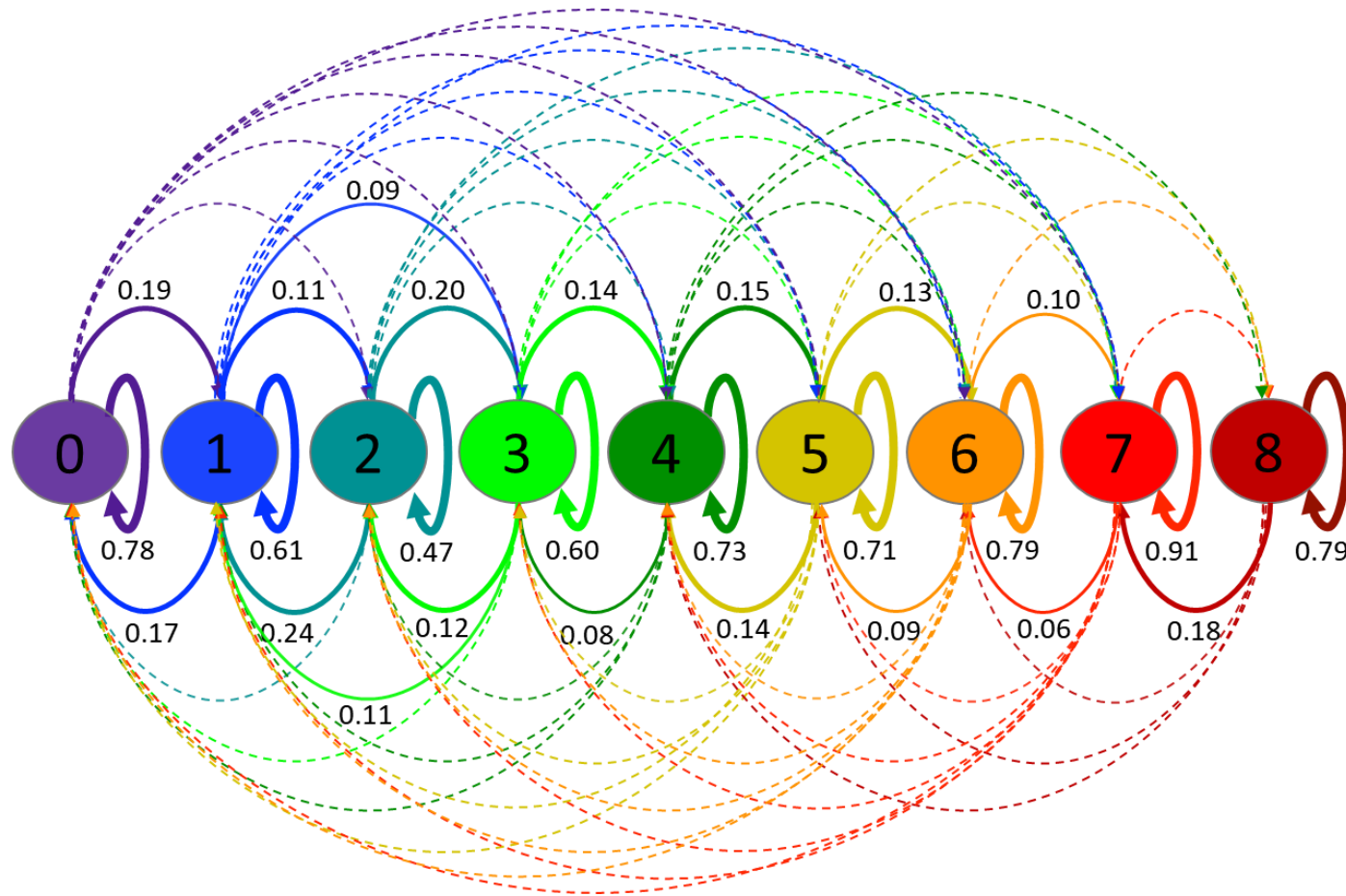
<i>Frequencies</i>									
	A	B	C	D	E	F	G	H	
A	2362	567	72	32	7	4	2	1	3047
B	529	1963	353	302	41	7	5	2	3202
C	58	354	693	291	65	6	6	1	1474
D	27	276	283	1478	349	26	12	1	2452
E	7	37	56	269	2506	515	52	10	3452
F	4	1	10	19	392	1973	352	27	2778
G	2	4	5	13	32	238	2157	290	2759
H	1	2	2	2	5	19	195	3127	3353
<i>Probabilities</i>									
	A	B	C	D	E	F	G	H	
A	0.78	0.19	0.02	0.01	0.002	0.001	0.001	0.0003	
B	0.17	0.61	0.11	0.09	0.01	0.002	0.002	0.001	
C	0.04	0.24	0.47	0.20	0.04	0.004	0.004	0.001	
D	0.01	0.11	0.12	0.60	0.14	0.01	0.005	0.0004	
E	0.002	0.01	0.02	0.08	0.73	0.15	0.02	0.003	
F	0.001	0.0004	0.004	0.01	0.14	0.71	0.13	0.01	
G	0.001	0.001	0.002	0.005	0.01	0.09	0.79	0.11	
H	0.0003	0.001	0.001	0.001	0.001	0.01	0.06	0.93	





**Table 11:** Scheme 2: Unadjusted lumped transition matrix, overall (N=2,047; 9 states to 5 lumps)

<i>Frequencies</i>						
	A	B	C	D	E	
A	2362	639	32	13	1	3047
B	587	3363	593	130	3	4676
C	27	559	1478	387	1	2452
D	13	113	301	8235	327	8989
E	1	4	2	219	3127	3353
<i>Probabilities</i>						
	A	B	C	D	E	
A	0.78	0.21	0.01	0.004	0.0003	
B	0.13	0.72	0.13	0.03	0.0006	
C	0.01	0.23	0.60	0.16	0.0004	
D	0.001	0.01	0.03	0.92	0.04	
E	0.0003	0.001	0.001	0.07	0.93	



**Figure 19:** State transition diagram for the overall sample, unlumped chain



	$p_{i,j} \geq 0.10$
	$0.10 < p_{i,j} < 0.25$
	$0.05 < p_{i,j} \leq 0.10$
	$p_{i,j} \leq 0.05$ ; values not shown

Arrows/connecting lines above the states represent transition to higher states; those below represent transitions to lower states or remaining in the same state. Probabilities approximately 0 are not depicted.

**Table 12:** Unadjusted transition matrices, Mild Impairment (N=1,545 for 16,995 transitions)

<i>Frequencies</i>	0	1	2	3	4	5	6	7	8
0	2261	542	71	30	7	3	2	0	0
1	505	1910	340	299	40	6	3	1	0
2	56	342	665	284	57	2	3	1	0
3	26	271	274	1461	333	23	11	0	0
4	6	36	49	253	2319	440	42	7	2
5	3	1	7	16	327	1373	247	16	1
6	1	3	2	9	20	162	1024	110	4
7	0	0	1	2	2	11	60	550	18
8	0	0	0	0	2	1	2	16	21

<i>Probabilities</i>	0	1	2	3	4	5	6	7	8
0	0.78	0.19	0.02	0.01	0.002	0.001	0.001	0	0
1	0.16	0.62	0.11	0.10	0.01	0.002	0.001	0.0003	0
2	0.04	0.24	0.47	0.20	0.04	0.001	0.002	0.001	0
3	0.01	0.11	0.11	0.61	0.14	0.01	0.005	0	0
4	0.002	0.01	0.02	0.08	0.74	0.014	0.01	0.002	0.001
5	0.002	0.001	0.004	0.01	0.16	0.69	0.12	0.01	0.001
6	0.001	0.002	0.001	0.01	0.01	0.12	0.77	0.08	0.003
7	0	0	0.002	0.003	0.003	0.02	0.09	0.85	0.03
8	0	0	0	0	0.05	0.02	0.05	0.38	0.50

**Table 13:** Scheme 1: Unadjusted, lumped transition matrix, Mild Impairment (N=1,545; 9 states to 8 lumps)

<i>Frequencies</i>	A	B	C	D	E	F	G	H
A	2261	542	71	30	7	3	2	0
B	505	1910	340	299	40	6	3	1
C	56	342	665	284	57	2	3	1
D	26	271	274	1461	333	23	11	0
E	6	36	49	253	2319	440	42	9
F	3	1	7	16	327	1373	247	17
G	1	3	2	9	20	162	1024	114
H	0	0	1	2	4	12	62	605

<i>Probabilities</i>	A	B	C	D	E	F	G	H
A	0.78	0.19	0.02	0.01	0.002	0.001	0.001	0
B	0.16	0.62	0.11	0.10	0.01	0.002	0.001	0.0003
C	0.04	0.24	0.47	0.20	0.04	0.001	0.002	0.001
D	0.01	0.11	0.11	0.61	0.14	0.01	0.005	0
E	0.002	0.01	0.02	0.08	0.74	0.14	0.01	0.003
F	0.002	0.001	0.004	0.01	0.16	0.69	0.12	0.01
G	0.001	0.002	0.001	0.01	0.01	0.12	0.77	0.09
H	0	0	0.001	0.003	0.01	0.02	0.09	0.88

**Table 14:** Scheme 2: Unadjusted, lumped transition matrix, Mild Impairment (N=1,545; 9 states to 5 lumps)

<i>Frequencies</i>		A	B	C	D	E	
	A	2261	613	30	12	0	2916
	B	561	3257	583	111	2	4514
	C	26	545	1467	367	0	2399
	D	10	98	278	5954	140	6480
	E	0	1	2	78	605	686
<i>Probabilities</i>		A	B	C	D	E	
	A	0.78	0.21	0.01	0.004	0	
	B	0.12	0.72	0.13	0.02	0.0004	
	C	0.01	0.23	0.61	0.15	0	
	D	0.002	0.02	0.04	0.92	0.02	
	E	0	0.001	0.003	0.11	0.88	

**Table 15:** Unadjusted transition matrices, High Impairment (N=502 for 5,522 transitions)

<i>Frequencies</i>		0	1	2	3	4	5	6	7	8	
	0	101	25	1	2	0	1	0	1	0	131
	1	24	53	13	3	1	1	2	1	0	98
	2	2	12	28	7	8	4	3	0	0	64
	3	1	5	9	17	16	3	1	1	0	53
	4	1	1	7	16	187	75	10	1	0	298
	5	1	0	3	3	65	600	105	10	0	787
	6	1	1	3	4	12	76	1151	174	2	1424
	7	1	2	1	0	1	7	132	2332	31	2507
	8	0	0	0	0	0	0	1	21	138	160
<i>Probabilities</i>		0	1	2	3	4	5	6	7	8	
	0	0.77	0.19	0.01	0.02	0	0.01	0	0.01	0	
	1	0.24	0.54	0.13	0.03	0.01	0.01	0.02	0.01	0	
	2	0.03	0.19	0.44	0.11	0.13	0.06	0.05	0	0	
	3	0.02	0.09	0.17	0.32	0.30	0.06	0.02	0.02	0	
	4	0.003	0.003	0.02	0.05	0.63	0.25	0.03	0.003	0	
	5	0.001	0	0.004	0.004	0.08	0.76	0.13	0.01	0	
	6	0.001	0.001	0.002	0.003	0.01	0.05	0.81	0.12	0.001	
	7	0.0004	0.001	0.0004	0	0.0004	0.003	0.05	0.93	0.01	
	8	0	0	0	0	0	0	0.01	0.13	0.86	

**Table 16** : Scheme 1: Unadjusted, lumped transition matrix, High Impairment (N=502; 9 states to 8 lumps)

---

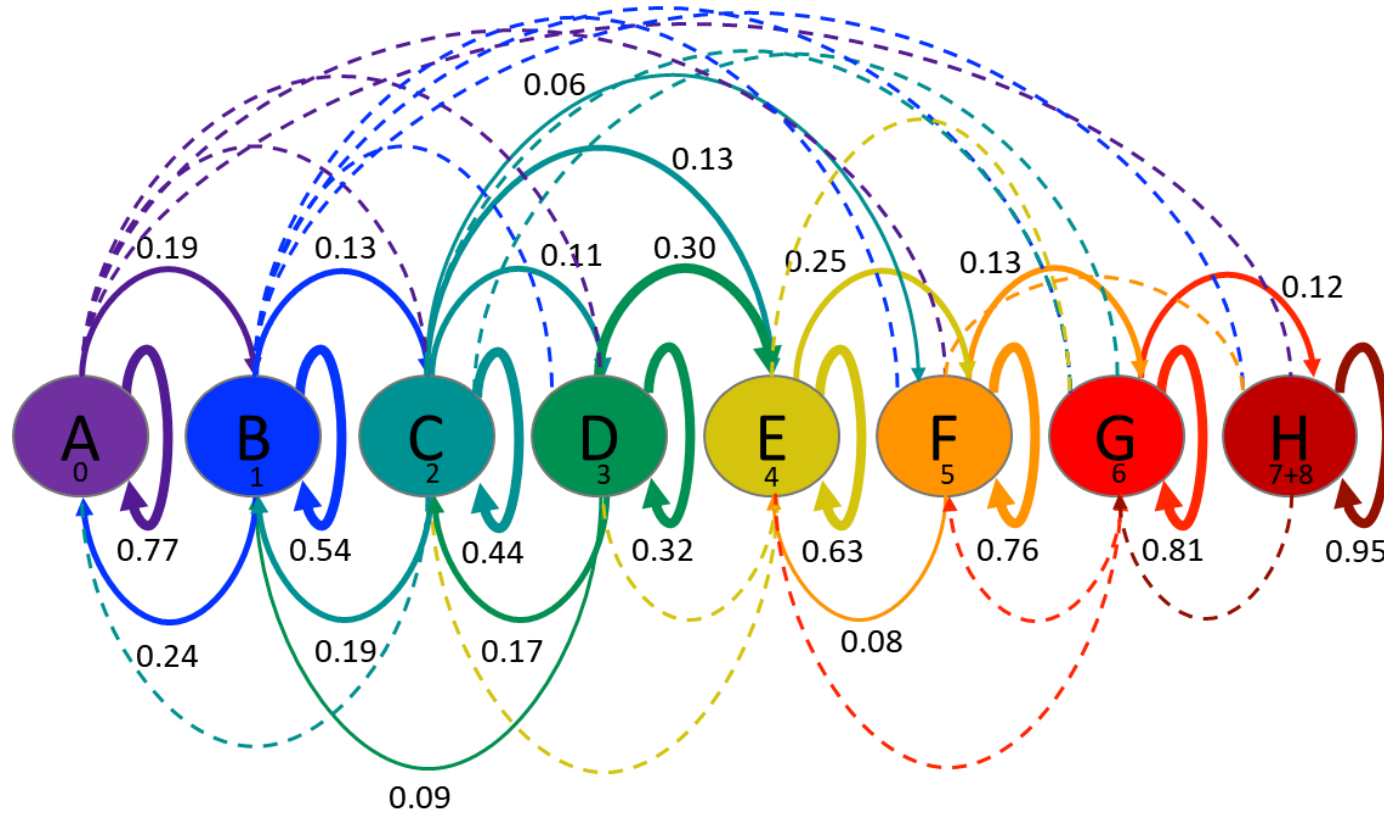
<i>Frequencies</i>									
	<i>A</i>	<i>B</i>	<i>C</i>	<i>D</i>	<i>E</i>	<i>F</i>	<i>G</i>	<i>H</i>	
<i>A</i>	101	25	1	2	0	1	0	1	131
<i>B</i>	24	53	13	3	1	1	2	1	98
<i>C</i>	2	12	28	7	8	4	3	0	64
<i>D</i>	1	5	9	17	16	3	1	1	53
<i>E</i>	1	1	7	16	187	75	10	1	298
<i>F</i>	1	0	3	3	65	600	105	10	787
<i>G</i>	1	1	3	4	12	76	1151	176	1424
<i>H</i>	1	2	1	0	1	7	133	2522	2667

<i>Probabilities</i>									
	<i>A</i>	<i>B</i>	<i>C</i>	<i>D</i>	<i>E</i>	<i>F</i>	<i>G</i>	<i>H</i>	
<i>A</i>	0.77	0.19	0.01	0.02	0	0.01	0	0.01	
<i>B</i>	0.24	0.54	0.13	0.03	0.01	0.01	0.02	0.01	
<i>C</i>	0.03	0.19	0.44	0.11	0.13	0.06	0.05	0	
<i>D</i>	0.02	0.09	0.17	0.32	0.30	0.06	0.02	0.02	
<i>E</i>	0.003	0.003	0.02	0.05	0.63	0.25	0.03	0.003	
<i>F</i>	0.001	0	0.004	0.004	0.08	0.76	0.13	0.01	
<i>G</i>	0.001	0.001	0.002	0.003	0.01	0.05	0.81	0.12	
<i>H</i>	0.0004	0.001	0.0004	0	0.0004	0.003	0.05	0.95	

---

**Figure 20:** State transition diagram for lumped chain according to Scheme 1, High Impairment



———  $p_{i,j} \geq 0.10$   
 ———  $0.10 < p_{i,j} < 0.25$   
 ———  $0.05 < p_{i,j} \leq 0.10$   
 - - - -  $p_{i,j} \leq 0.05$ ; values not shown

Arrows/connecting lines above the states represent transition to higher states; those below represent transitions to lower states or remaining in the same state. Probabilities approximately 0 are not depicted.

**Table 17:** Scheme 2 : Unadjusted, lumped transition matrix, High Impairment (N=502; 9 states to 5 lumps)

<i>Frequencies</i>	A	B	C	D	E	
A	101	26	2	1	1	131
B	26	106	10	19	1	162
C	1	14	17	20	1	53
D	3	15	23	2281	187	2509
E	1	3	0	141	2522	2667
<i>Probabilities</i>	A	B	C	D	E	
A	0.77	0.20	0.02	0.01	0.01	
B	0.16	0.65	0.06	0.12	0.01	
C	0.02	0.26	0.32	0.38	0.02	
D	0.001	0.01	0.01	0.91	0.07	
E	0.0004	0.001	0	0.05	0.95	

**Table 18:** Summary of results for complete case and stratified by enrollment disability, unadjusted probabilities

Sample	Scheme	States	Lumps	$\chi^2_{\dagger}$	DF $_{\ddagger}$	p-value
<i>Overall</i>	1	9	8	17.92	7	0.0123
<i>(N=2,047)</i>	2	9	5	1223.83	16	<0.0001
<i>Mildly Impaired</i>	1	9	8	14.60	5	0.0122
<i>(N=1,545)</i>	2	9	5	819.66	15	<0.0001
<i>Highly Impaired</i>	1	9	8	7.68	6	<b>0.2628</b>
<i>(N=502)</i>	2	9	5	248.24	15	<0.0001

$\dagger$ Chi-square test of lumpability test statistic

$\ddagger$ Degrees of Freedom

### 3. Secondary Analyses

**Overall sample** For the overall, unlumped sample, only current PDDS and enrollment PDDS and race are associated with the transitions ( $p < 0.05$ ); no other predictors were associated with the transitions (Table 19). Both PDDS predictors

has negative parameter estimates, therefore indicating a cumulative probability of higher scores (increased disability). For both Scheme 1 and Scheme 2, enrollment PDDS, current PDDS and race were associated with the transitions ( $p < 0.05$ ); no other predictors were associated with the transitions (Table 19). The parameter estimates for both schemes are consistent with the overall unlumped model: the parameter estimates are negative and therefore indicate a cumulative probability of higher PDDS scores.

**Table 19:** Summary of proportional odds models, overall sample<sup>†</sup>

<b>Lumping scheme</b>	<b>Predictor</b>	<b>Estimate (SE<sub>‡</sub>)</b>	<b>p-value*</b>
<i>None</i>	Current PPDS (lumped)	-2.51 (0.05)	<b>&lt;0.0001</b>
	Enrollment PDDS	-0.23 (0.01)	<b>&lt;0.0001</b>
	Gender (Female)	0.10 (0.03)	<b>0.0017</b>
	Race (African American)	0.27 (0.11)	<b>0.0219</b>
	Race (Other)	0.10 (0.07)	
	Age of MS Diagnosis	0.002 (0.001)	0.1123
	Relapse (No)	-0.06 (0.05)	0.2617
	Relapse (Unsure)	-0.09 (0.07)	
<i>Scheme 1</i>	Current PPDS (lumped)	-2.40 (0.05)	<b>&lt;0.0001</b>
	Enrollment PDDS	-0.21 (0.02)	<b>&lt;0.0001</b>
	Gender (Female)	0.09 (0.03)	<b>0.0074</b>
	Race (African American)	0.21 (0.13)	0.1613
	Race (Other)	0.08 (0.08)	
	Age of MS Diagnosis	0.001 (0.002)	0.4851
	Relapse (No)	-0.05 (0.06)	0.4540
	Relapse (Unsure)	-0.07 (0.07)	
<i>Scheme 2</i>	Current PPDS (lumped)	-1.08 (0.05)	<b>&lt;0.0001</b>
	Enrollment PDDS	-1.58 (0.07)	<b>&lt;0.0001</b>
	Gender (Female)	0.51 (0.15)	<b>0.0006</b>
	Race (African American)	0.77 (0.54)	0.1662
	Race (Other)	0.48 (0.37)	
	Age of MS Diagnosis	0.01 (0.01)	0.1555
	Relapse (No)	-0.11 (0.23)	0.5923
	Relapse (Unsure)	-0.31 (0.33)	

<sup>†</sup> Random effects (random intercept) model

<sup>‡</sup>Standard Error

\*Type III Test of Fixed Effects

**Stratified by enrollment disability: Mildly Impaired** The proportional odds model with random intercepts was fit for the unlumped MI group (Table 20). When fitting the random intercept proportional odds models to the lumped MI group, we found estimation of the model parameters to be numerically unstable and the models would not converge. Thus the stratified samples with lumping were modeled using proportional odds models without random intercepts (fixed effects). In all cases current and enrollment PDDS were statistically associated with transitions, such that higher predictive PDDS scores are associated with an increased probability of higher “next” PDDS scores. Gender was only associated with the transitions when implementing Scheme 2. Race, relapse history and age of diagnosis were not significantly associated with score transitions in any of the stratified models (all  $p > 0.05$ ).

**Stratified by enrollment disability: Highly Impaired** The proportional odds model with random intercepts was fit for the unlumped HI group (Table 21). When fitting the random intercept proportional odds models to the lumped HI group, we found estimation of the model parameters to be numerically unstable and the models would not converge. Thus the stratified samples with lumping were modeled using proportional odds models without random intercepts (fixed effects). In all cases current and enrollment PDDS were statistically associated with transitions, such that higher predictive PDDS scores are associated with an increased probability of higher “next” PDDS scores. Race was only associated with the transitions in the unlumped model. Gender was only associated with transitions in the unlumped model and under Scheme 2; therefore, being female is associated



with having lower PDDS scores. Relapse history and age of diagnosis were not significantly associated with score transitions in any of the stratified models (all  $p>0.05$ ).

**Table 20:** Summary of proportional odds models, Mildly Impaired

<b>Lumping scheme</b>	<b>Predictor</b>	<b>Estimate (SE<sub>1</sub>)</b>	<b>p-value<sub>2</sub></b>
<i>None</i> <sup>†</sup>	Current PPDS	-2.39 (0.05)	<b>&lt;0.0001</b>
	Enrollment PDDS	-0.19 (0.02)	<b>&lt;0.0001</b>
	Gender (Female)	0.06 (0.03)	0.0801
	Race (African American)	0.15 (0.12)	0.4482
	Race (Other)	-0.02 (0.08)	
	Age of MS Diagnosis	-0.001 (0.002)	0.4497
	Relapse (No)	-0.01 (0.05)	
	Relapse (Unsure)	-0.9 (0.08)	0.4596
<i>Scheme 1</i> <sup>‡</sup>	Current PPDS	-2.34 (0.02)	<b>&lt;0.0001</b>
	Enrollment PDDS	-0.21 (0.01)	<b>&lt;0.0001</b>
	Gender (Female)	0.07 (0.04)	0.0802
	Race (African American)	0.11 (0.15)	0.7588
	Race (Other)	-0.01 (0.09)	
	Age of MS Diagnosis	0.002 (0.002)	0.3077
	Relapse (No)	-0.02 (0.05)	
	Relapse (Unsure)	-0.09 (0.07)	0.4234
<i>Scheme 2</i> <sup>‡</sup>	Current PPDS (lumped)	-1.92 (0.02)	<b>&lt;0.0001</b>
	Enrollment PDDS	-0.29 (0.02)	<b>&lt;0.0001</b>
	Gender (Female)	0.13 (0.05)	<b>0.0092</b>
	Race (African American)	0.06 (0.21)	0.9602
	Race (Other)	0.002 (0.11)	
	Age of MS Diagnosis	-0.0004 (0.002)	0.8539
	Relapse (No)	0.30 (0.70)	
	Relapse (Unsure)	-0.07 (0.09)	0.6789

<sup>†</sup> Random effects (random intercept) model

<sup>‡</sup> Fixed effects model

<sub>1</sub>Standard Error

<sub>2</sub> Type III Test of Fixed Effects

**Table 21:** Summary of proportional odds models, Highly Impaired

<b>Lumping scheme</b>	<b>Predictor</b>	<b>Estimate (SE<sub>1</sub>)</b>	<b>p-value<sub>2</sub></b>
<i>None</i> <sup>†</sup>	Current PPDS	-2.73 (0.23)	<b>&lt;0.0001</b>
	Enrollment PDDS	-1.32 (0.21)	<b>&lt;0.0001</b>
	Gender (Female)	0.37 (0.17)	<b>0.0327</b>
	Race (African American)	0.77 (0.31)	<b>0.0119</b>
	Race (Other)	0.77 (0.38)	
	Age of MS Diagnosis	0.01 (0.01)	0.3451
	Relapse (No)	-0.38 (0.23)	0.2389
	Relapse (Unsure)	-0.14 (0.29)	
<i>Scheme 1</i> <sup>‡</sup>	Current PPDS	-2.75 (0.05)	<b>&lt;0.0001</b>
	Enrollment PDDS	-0.65 (0.06)	<b>&lt;0.0001</b>
	Gender	0.16 (0.08)	0.0529
	Race (African American)	0.33 (0.25)	0.0560
	Race (Other)	0.34 (0.16)	
	Age of MS Diagnosis	0.004 (0.004)	0.2800
	Relapse (No)	-0.19(0.11)	0.2298
	Relapse (Unsure)	-0.03 (0.14)	
<i>Scheme 2</i> <sup>‡</sup>	Current PPDS (lumped)	-1.89 (0.04)	<b>&lt;0.0001</b>
	Enrollment PDDS	-0.75 (0.08)	<b>&lt;0.0001</b>
	Gender (Female)	0.38 (0.11)	<b>0.0007</b>
	Race (African American)	0.49 (0.31)	0.1920
	Race (Other)	0.23 (0.22)	
	Age of MS Diagnosis	0.004 (0.005)	0.4429
	Relapse (No)	-0.18 (0.14)	0.4246
	Relapse (Unsure)	0.06 (0.20)	

<sup>†</sup> Random effects (random intercept) model

<sup>‡</sup> Fixed effects model

<sup>1</sup>Standard Error

<sup>2</sup> Type III Test of Fixed Effects

**Summary** These models consistently had negative parameter estimates for current PDDS and enrollment PDDS; therefore indicating a cumulative probability of higher scores (increased disability) with increasing current and enrollment PDDS. Race and gender were not consistently associated with transitions. Where gender was statistically significant, the nature of the relationship was consistent: being female is

associated with an increased probability of having lower PDDS scores. For all models using the overall sample, gender is associated with the outcome; the HI sample did not have a consistent statistical association with gender. For both disability groups, implementation of Scheme 1 did not yield an association with gender, but the relationship is present when implementing Scheme 2.

#### D. Discussion and Conclusion

This chapter explored the application of the Test of Lumpability and our novel, proposed Goodness of Fit test on patient reported disability scores (the PDDS) obtained from a registry. We demonstrated the challenge involved with identifying a simultaneous parsimonious and useful lumping scheme that also retains the Markov property. In doing so, we further demonstrated that even with retaining this property, lumping by a specific scheme may not be the best fit to the data.

*Primary results* In the primary analyses, all transition matrices share a common characteristic of transitions being concentrated on the diagonal, confirming the inherent dependency on previous scores and that knowledge of previous disability status is highly informative for considering future disability. In the matrix representing the overall sample, the distribution of transitions along the diagonal are relatively evenly distributed between the PDDS scores. The distribution of transitions in the stratified samples are concentrated about those PDDS scores that drove the disability classification at enrollment. That is, the matrix for MI sees most

of the transitions occur among scores ranging 0 to 6; in the HI matrix, the transitions are concentrated in among scores 5-8. Interestingly, there are nonzero transitions at the lower scores (0-4) in the HI matrix. While unusual, as the grouping is based on enrollment scores, there is the possibility that the enrollment scores were recorded while the patients was undergoing a relapse, their experience of using the PDDS improved over time, or reporting or recording error.

However, in performing the Test of Lumpability, we observed that none of the chains were lumpable at the conventional 0.05 level for any other than the HI group under Scheme 1. As it only experienced a combination of the highest 2 states (PDDS scores 7 and 8), these transitions are consistent with the observed states in the HI's unlumped chain until state H (combination of 7 and 8). Examination of the state diagram (Figure) demonstrates the complexity still present in the chain, and reinforces the motivation to identify useful lumping scheme(s). The fact that this chain is for a subset of the NARCOMS sample suggests that such a scheme or schemes will need to be specific for certain disease stages in order for them to be accurate and useful. Because disease worsens with time, and worsening disease (decreased mobility) is described by higher PDDS scores, these results support the concept that different stages of the disease should be treated differently. Due to only a single scheme for a single subset being able to pass the test of lumpability, additional work to identify other scientifically supported and clinically meaningful schemes.

In performing the Test of Lumpability, we observed that most of the chains were not lumpable at the conventional 0.05 level (other than the HI group under Scheme 1). It is reasonable that with a large enough sample, there is not a great need to lump. Therefore it is not surprising that we fail the test; the results do lead us to our future work. Because we observed the test of lumpability seems to be influenced by sample size, it seems further adjustment to the test is necessary so that the results are not driven by sample size. Returning to the significance level, the larger sample size suggests we can be stricter about significance level, but more work needs to be done to develop guidelines on how to adjust an  $\alpha$ -level according to sample size. By extension, perhaps the Test of Lumpability is not adequately penalizing for the number of parameters, because as the sample increases, we are less likely to choose the lumped chain. However, having a large sample is not a drawback, as it provides framework to be more specific in the model regarding number of states and lumps, and also the benefit of using covariates to obtained adjusted transition probabilities. Additionally, because it is implicit in the Test of Lumpability that the original chain and the lumped chain must share the same order (first order), there is the possibility that the lumped chain is larger than order 1.

*Secondary results* The secondary results were for covariate-adjusted matrices.

Because of the size of the dataset and the complexity of the model, we did have to determine which method would lead to model convergence. While slower, Newton-Raphson was selected due to its reliability. For several of the models, SAS reported a log warning that “at least one element of the gradient is greater than 1e-3;” this warning was observed in the presence of random effects, but not when random

effects were removed. This warning typically means we might have larger standard errors about our estimates; removal of random effects produced results highly similar to those observed with them, supporting continued use of random effects in the model. Taken together with the value of accounting for disease heterogeneity via random effects, we proceeded in the presence of this message. Other random effects were considered (namely, current PDDS); however, this introduced additional complicity to the model that generated convergence issues. The fact we could not employ random effects models in certain settings demonstrates the complexity already present in the dataset and in the model.

We observed largely consistent effects for those PDDS predictors with a statistically significant association with the outcome. Of particular note is that in all model-lumping scheme scenarios, the current and enrollment PDDS were associated with movement between states. This provides further evidence supporting the nature of the disability changes is Markovian.

Only the Overall sample has consistent predictor associations; upon stratifying, these associations varied between strata and lumping schemes. Interestingly, the association with gender varied according to lumping scheme and was consistent between the two strata. Race was only associated with transitions for the overall, unlumped sample; because the study sample is overwhelmingly Caucasian, any differences due to race would have been difficult to detect. Gender differences were not observed for Scheme 1 (9 states to 8 lumps), but these differences are present under application of Scheme 2 (9 states to 5 lumps); this is counterintuitive, as

Scheme 1 has greatest resemblance to the original, unlumped chain compared to Scheme 2 and the unlumped chains. A potential explanation is less sparseness in the transition matrix is observed for a chain with fewer states (Scheme 2) and therefore less variability between probabilities. Additionally, it suggests further exploration is needed to tease out the influence of Gender.

As one of our aims was to examine the relationship between relapse and transitions between scores, we were surprised that relapse history was not a significant predictor in the model. It is possible that more specific information regarding relapse history might be useful (such as treatment details or disease duration), or else there is an interaction term involving relapse that might provide insight into the relationship.

We did anticipate a problem with continuity of response and so although we began with an initial dataset that was a large, it was substantially restricted due to missing responses, a known issue encounter with self-report data collection. Because of the inclusion criterion of having complete PDDS data, there is a potential for selection bias in our sample.

It was of interest to evaluate the proportional odds assumption, particularly for the Current PDDS predictor. However, due model complexity and the multitude of additional steps to resolve, this is an additional avenue of future work. While we successfully modeled transitions using our chosen predictors, we did not perform tests on adjusted probability matrices. This is a logical step to extend these analyses and will require different transition matrices for every subject at every time point in

order to obtain likelihood estimates, and ultimately perform the test of lumpability to determine if controlling for certain patient characteristics affects the lumpability of the Markov process.

The primary results were for unadjusted matrices, and consideration of other lumping schemes are of interest are to be explored. As mentioned earlier, we assumed time homogeneity because of the short period of observation, relative to lifetime disease duration. However, because the disease changes with time, albeit very slowly, it seems prudent for further investigations of the chains with consideration to disease duration. This includes more extensive disease duration time periods (evaluation of stationarity), model order, whether certain schemes are more appropriate for specific disease duration. Because of the sample size and long-term nature of the registry, there is great potential to tease out where and how stationarity might change. That is, where it might be appropriate to have a separate chain based on time periods of disease duration and even disability level. Finally, it is of interest to apply this methodology to EDSS and compare identify common lumping schemes, GOF performance, and performance of proportional odds model and associated predictors.



## ***V. Applications to the CombiRx Trial Data: EDSS and Lumpability***

### **A. Introduction**

The overarching goal of this section is to investigate disability progression using Markovian methodology to model the expanded disability status scale (EDSS) and employ covariates which contribute to the movement between EDSS scores, utilizing data from the Combination Therapy in Patients With Relapsing-Remitting Multiple Sclerosis (CombiRx) trial (N=725). Our specific goals were to (1) identify and implement useful lumping schemes for EDSS; (2) demonstrate the applicability of the Test of Lumpability for the EDSS; and (3) implement the 2-stage lumpability assessment process. While the EDSS has been modeled using Markov chains, implementation of the Test of Lumpability will be novel for this outcome (Engler et al., 2017)(Healy & Engler, 2009). Additionally, analysis of the CombiRx data utilizing Markov methodology will also be novel.

### **B. Methods**

#### **1. Study Design**

*Study Design* CombiRx was a randomized, placebo-controlled, 3-arm, double-blind multi-site clinical trial (Phase III) (Lindsey et al., 2012). Enrollment EDSS score less

than 5.5 was the cutoff for trial participation. Patients were randomized to either Glatiramer acetate (GA) Interferon- $\beta$  (IFN) or combination (IFN+GA). Half of the patients were randomized to the combination therapy, with 25% in each of the remaining therapeutic arms. Study duration was 36 months, with follow-up occurring every 3 months. However, only the 6-month follow-up intervals were included in these analyses, so as to maintain ability for comparison between these results and those NARCOMS's (previous chapter).

*Protocol approval and patient consent* All data were de-identified and participants gave informed consent. This study was approved by the institutional review board at the University of Alabama at Birmingham, Birmingham, Alabama.

This is a longitudinal interventional study. The primary variables of interest is EDSS at enrollment and at follow-up. Other enrollment variables considered included gender (male, female), race (African American, Caucasian, Other), body mass index (BMI), year of enrollment, age at baseline (years), year of MS diagnosis, age at MS diagnosis (years), time since symptom onset (years), disease duration at enrollment (years), number of relapses (in the previous 12 months and previous 3 years), treatment group (GA, IFN, IFN+GA), and marital status (divorced, married/cohabitating, single, separated) (Table 1). Follow-up data included in the analyses was EDSS only for every 6-month follow-up through 36 months, resulting in 6 time points.

*Participants* Participants included adult male and female MS patients participating in CombiRx (aged 18-60 years). Of the 1,008 participants enrolled, 978 had at least

one update in the 3-year observation period; of those 725 had complete follow-up responses for EDSS. Therefore, only patients with complete-case follow-up responses for EDSS were included in the analytic cohort (N=725).

*Inclusion criteria* Participants with baseline enrollment information and 6-month EDSS updates for the duration of the 3-year study period were considered for analysis.

*Exclusion criteria* Participants with any missing EDSS responses (baseline and at 6-month follow-up intervals) were excluded from analysis.

*Preliminary Analysis* Data were summarized using means, standard deviations (SD), frequencies (N), medians, minima and maxima (Table 22).

## 2. Primary Analyses

The primary outcome is the transition of EDSS scores, which is considered at the 6-month follow-up points for the 3-year duration of the regular study period. The original chain has 20 states (one state per EDSS score). The appropriate transition matrices are presented for the overall sample, without stratification and unadjusted for any covariates (Tables 23). Due to the short study period relative to expected disease duration, the analyses were performed under the assumption of time homogeneity. Three lumping schemes were considered to reduce the number of states from the original chain (Figure 20). Each scheme was implemented and Test

of Lumpability was performed on the pair of unlumped and lumped chains (Jernigan & Baran, 2003).

While the unlumped chain should contain 20 states (Figure 20), based on the total number of EDSS scores observed, the unlumped chain in this case will be a 14-state chain, to reflect the maximum score observed in the study population. Therefore, the unlumped chain is designated as *Scheme 0* and contains 14 states. *Scheme 1*, named “Baseline groupings,” reflects the CombiRx groupings for baseline EDSS, resulting in 4 lumps from 14 states (G. Wang et al., 2017). This lumping scheme is also supported by results produced by Engler et al. (2017). *Scheme 2*, named “Simple Combination,” reduces the number of observed states to half, by combining every 2 contiguous states. While again noting there is no one-to-one correspondence between EDSS and PDDS, *Scheme 3*, named “PDDS matching,” reflects their similarity (Hohol et al., 1995; Hohol et al., 1999; Learmonth et al., 2013; Marrie et al., 2006; Marrie & Goldman, 2007). This lumping scheme is therefore similar to the method employed with aggregating the PDDS to reflect the correspondence between PDDS and EDSS (Figure 20-21).

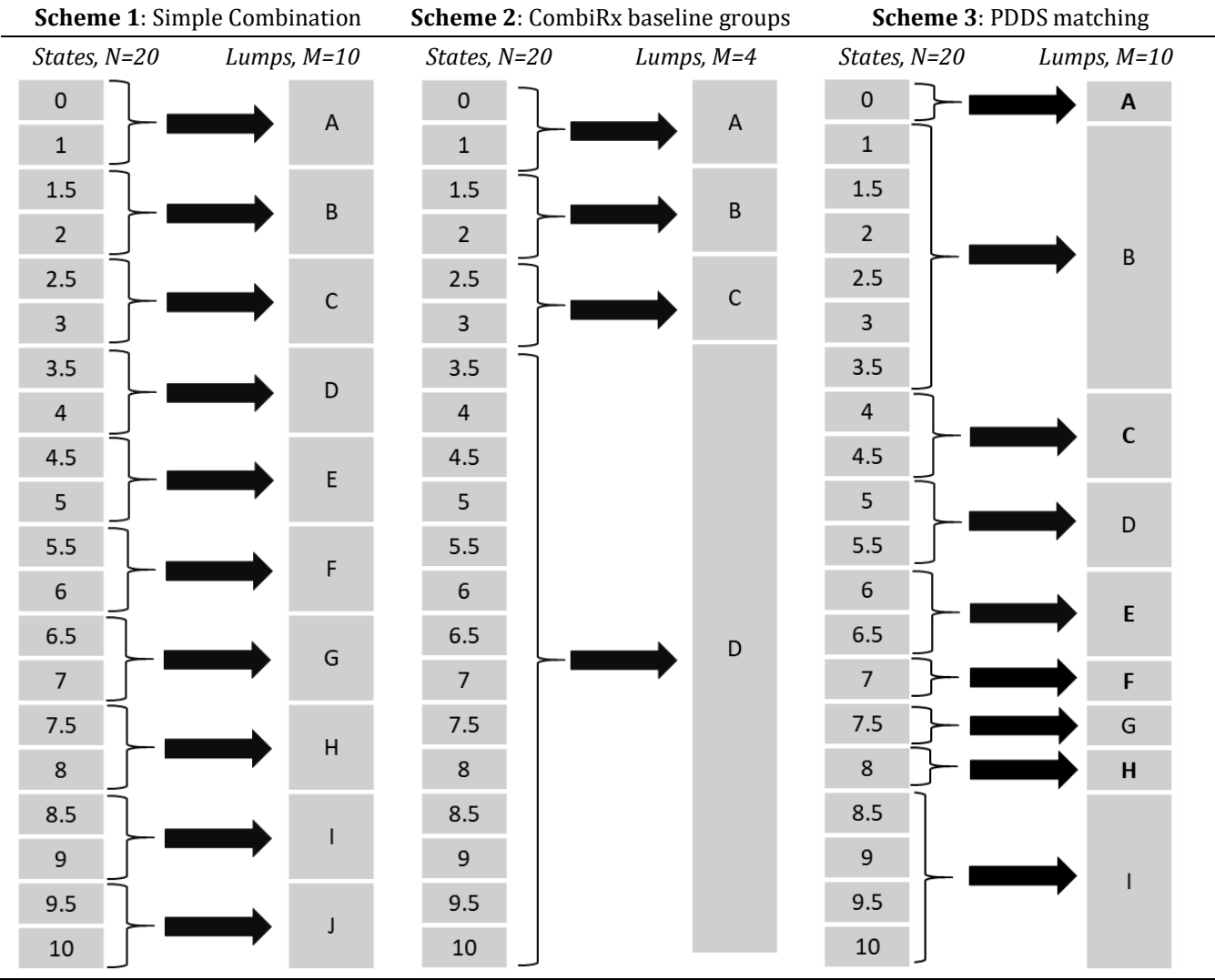
The proposed Chi-square goodness of fit tests comparing the lumped and unlumped chains were performed on the schemes with  $p \geq 0.05$  for the test of lumpability.

### 3. Secondary Analyses

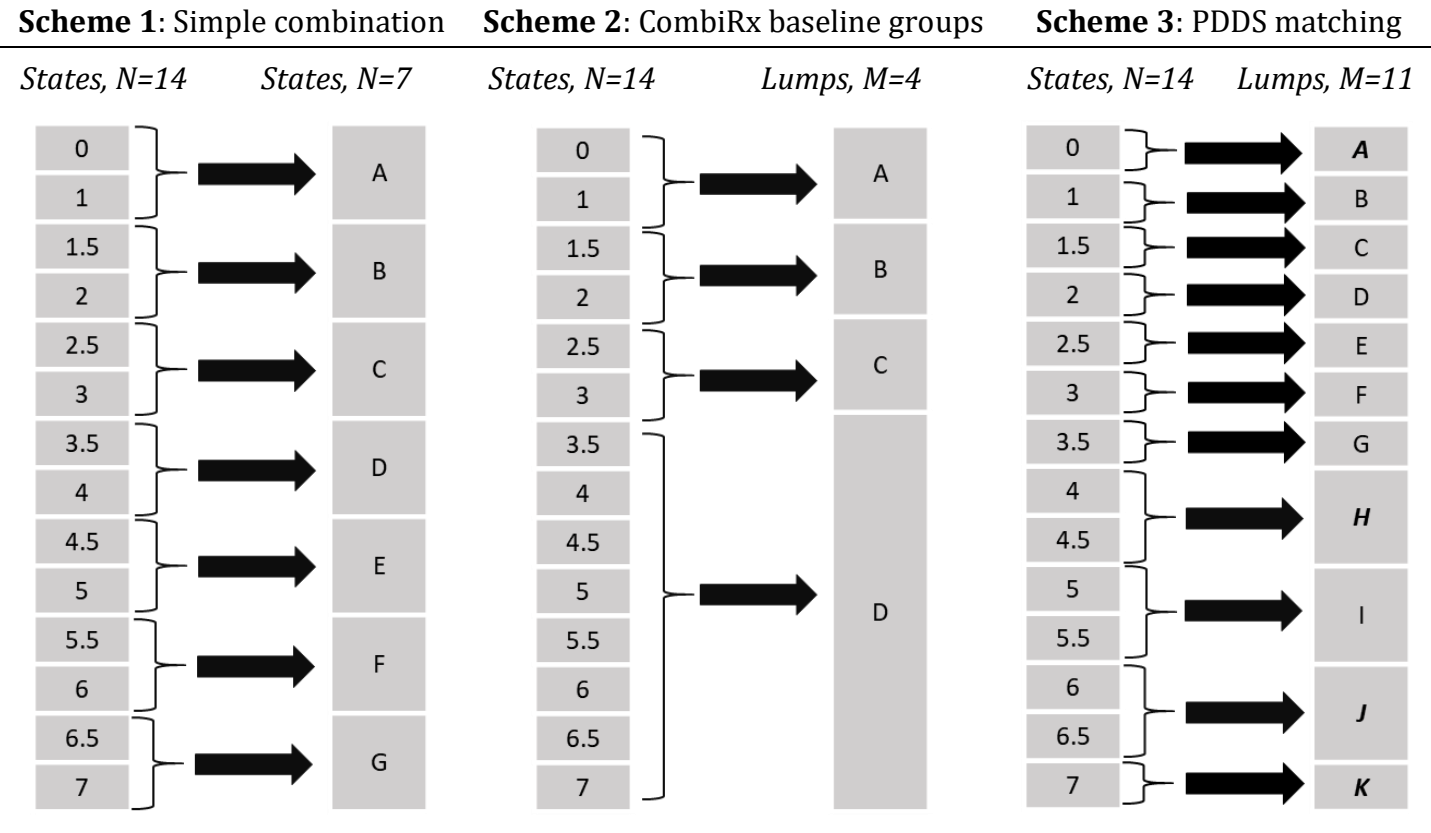
Transitions were modeled, adjusted for covariates using proportional odds modeling with random effects (intercept), for the overall sample. Random effects were included to account for the heterogeneous nature of the disease between patients. Optimization was performed via the Newton-Raphson method, and Compound Symmetry was used as the covariance structure. Predictors included gender, race, age of diagnosis, EDSS at baseline, and current EDSS to predict the next EDSS. These variables are consistent with those utilized in modeling PDDS, again for the purposes of comparison (refer to previous chapter). For numeric stability and model convergence, Current EDSS and Baseline EDSS were treated as numeric. When using these lumped “current” EDSS scores, the mean EDSS score was used for prediction. Specifically, for Scheme 1 (Simple combination), Lump A’s numeric value is 0.5, as it is the mean for 0 and 1. Lump B’s numeric value was 1.75 (since 1.5 and 2 are combined); Lump C is 2.75, and so forth.

Statistical analyses were performed in SAS V9.4, SAS/IML v14.3 and JMP Pro V14.0 (SAS Institute, Inc., Cary, NC) under an  $\alpha=0.05$  significance level, unless otherwise indicated.

Figure 21: Proposed lumping schemes for all EDSS scores



**Figure 22 :** Lumping schemes as applied to the EDSS scores from study period



## C. Results

### 1. Preliminary Analyses

Although not numerically the same, the summary statistics of the analytic cohort's baseline characteristics are consistent with those reported at baseline for the entire randomized CombiRx cohort (Tables 22, 23) (Lindsey et al., 2012). There were patients with 2 protocol violations at baseline (EDSS=6 and 6.5) (Lindsey et al., 2012); one of these patients is included in the analytic cohort (EDSS=6; Table 23). All participants had at least one relapse at baseline.

**Table 22:** Summary of cohort demographics at enrollment or baseline (N=725)

<i>Demographics and diagnosis</i>	<b>N</b>	<b>%</b>
<b>Gender</b>		
Female	521	71.9
Male	204	28.1
<b>Race</b>		
African American	48	6.6
Caucasian	642	88.5
Other	35	4.8
<b>Marital status</b>		
Divorced	59	8.1
Married/cohabitating	451	62.2
Single	204	28.1
Separated	11	1.5
<b>Year of Enrollment</b>		
2005	120	16.5
2006	268	37.1
2007	182	25.1
2008	118	16.3
2009	36	5.0
	<b>Mean (SD<sup>†</sup>)</b>	<b>Median (Min, Max)</b>
<b>Age at baseline</b> (years)	38.4 (9.4)	38.0 (18.0, 61.0)
<b>Age at diagnosis</b> (years)	39.6 (10.4)	39.0 (18.0, 78.0)
<b>Body Mass Index</b>	28.8 (6.8)	27.5 (16.1, 58.5)

<sup>†</sup> Standard deviation



**Table 23:** Summary of cohort clinical characteristics baseline (N=725)

<i>Baseline disease details</i>	<b>N</b>	<b>%</b>
<b>Treatment Group</b>		
GA	204	28.1
IFN	166	22.9
IFN + GA	355	49.0
<b>EDSS</b>		
0	95	13.1
1	99	13.7
1.5	107	14.8
2	179	24.7
2.5	91	12.5
3	61	8.4
3.5	52	7.2
4	25	3.4
4.5	5	0.7
5	4	0.5
5.5	6	0.8
6	1	0.1
	<b>Mean (SD<sup>†</sup>)</b>	<b>Median (Min, Max)</b>
<b>Disease Duration (years)</b>	1.1 (3.0)	0 (0, 23.0)
<b>Time since first symptom (years)</b>	4.3 (5.5)	2.0 (0, 39.0)
<b>Relapses in previous 12 months</b>	1.7 (0.8)	2.0 (0, 6.0)
<b>Relapses in previous 3 years</b>	2.4 (0.9)	2.0 (1.0, 10.0)
<b>EDSS (Numeric)</b>	1.9 (1.2)	2.0 (0, 6.0)

<sup>†</sup>Standard deviation

## 2. Primary Analyses

Without accounting for baseline EDSS or other covariates, the frequency and probability transition matrices depict relative consistency of the EDSS score over time. Most values are concentrated about the diagonal of the matrix (Tables 25-29). The 725 patients account for a total of 3,625 transitions across 6 time points.

The unlumped chain was lumpable only according to *Scheme 3* (PDDS Matching); it was not lumpable according to the other proposed schemes (Table 24). According to the Chi-square goodness of fit test, this lumping scheme produced a chain that was not a better fit to the data, compared to the unlumped chain. The conclusions between the LRT and Pearson formulations of the test are consistent (LRT:  $\chi^2(df = 72) = 235.37, p < 0.0001$ ; Pearson:  $\chi^2(df = 72) = 1907.80, p < 0.0001$ ) (Tables 24, 29; Figure 23).

**Table 24:** Summary of results for unadjusted matrices, N=725 ( $\alpha=0.05$ )

Scheme	N-states	M-lumps	$\chi^2_{\dagger}$	Degrees of Freedom	p-value
1	14	7	155.82	32	0.0
2	14	4	172.80	30	0.0
3	14	11	32.55	21	<b>0.0514</b>

$\dagger$ Chi-square test of lumpability test statistic

**Table 25:** Unadjusted, overall frequency transition matrix (N=725)

	0	1	1.5	2	2.5	3	3.5	4	4.5	5	5.5	6	6.5	7
0	237	119	46	55	15	5	3	1	1	0	0	0	0	482
1	114	205	96	102	21	7	4	1	0	0	0	0	0	550
1.5	47	103	154	133	47	13	10	0	1	0	1	0	1	510
2	54	109	132	343	131	73	22	9	0	0	1	0	1	874
2.5	19	23	46	139	114	55	38	12	2	1	0	2	0	451
3	6	10	13	52	67	93	46	20	1	2	2	4	0	316
3.5	4	1	4	30	25	44	60	34	3	0	3	2	3	213
4	1	0	3	9	7	18	33	44	1	4	2	4	3	129
4.5	0	0	1	1	0	2	2	3	7	0	4	4	0	24
5	0	0	1	0	0	1	2	4	2	0	0	1	0	11
5.5	0	0	1	1	0	0	2	2	2	3	7	3	0	21
6	0	0	0	2	1	0	3	2	2	0	3	16	2	31
6.5	0	0	1	0	0	0	1	1	0	0	0	0	8	12
7	0	0	0	0	0	0	0	0	0	0	0	0	1	1

**Table 26:** Unadjusted, overall probability transition matrix (N=725)

[illegible]

**Table 27:** Scheme 1: Simple combination, unadjusted transitions (14 states to 7 lumps)

<i>Frequencies</i>		<i>A</i>	<i>B</i>	<i>C</i>	<i>D</i>	<i>E</i>	<i>F</i>	<i>G</i>	
	<i>A</i>	675	299	48	9	1	0	0	1032
	<i>B</i>	313	762	264	41	1	1	2	1384
	<i>C</i>	58	250	329	116	6	8	0	767
	<i>D</i>	6	46	94	171	8	11	6	342
	<i>E</i>	0	3	3	11	9	9	0	35
	<i>F</i>	0	4	1	9	7	29	2	52
	<i>G</i>	0	1	0	2	0	0	10	13

<i>Probabilities</i>		<i>A</i>	<i>B</i>	<i>C</i>	<i>D</i>	<i>E</i>	<i>F</i>	<i>G</i>	
	<i>A</i>	0.62	0.29	0.05	0.01	0.001	0	0	
	<i>B</i>	0.23	0.55	0.19	0.03	0.001	0.001	0.001	
	<i>C</i>	0.07	0.32	0.43	0.15	0.01	0.01	0	
	<i>D</i>	0.02	0.13	0.27	0.50	0.02	0.03	0.02	
	<i>E</i>	0	0.08	0.08	0.31	0.26	0.26	0	
	<i>F</i>	0	0.08	0.02	0.17	0.13	0.56	0.04	
	<i>G</i>	0	0.08	0	0.15	0	0	0.77	

**Table 28:** Scheme 2: Baseline groupings unadjusted transitions (14 states to 4 lumps)

<i>Frequencies</i>		<i>A</i>	<i>B</i>	<i>C</i>	<i>D</i>	
	<i>A</i>	675	299	48	10	1032
	<i>B</i>	313	762	264	45	1384
	<i>C</i>	58	250	329	130	767
	<i>D</i>	6	54	98	284	442

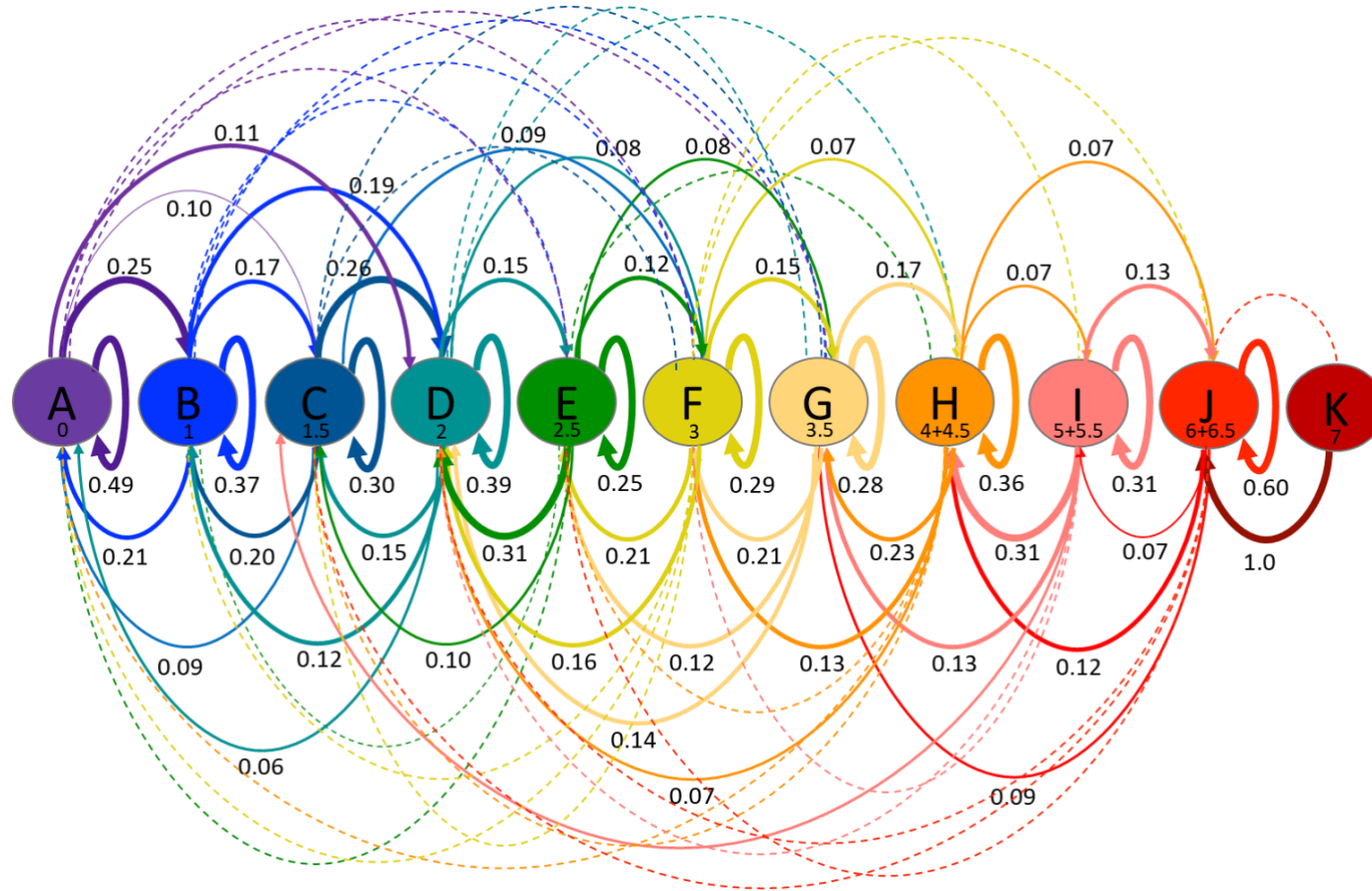
  

<i>Probabilities</i>		<i>A</i>	<i>B</i>	<i>C</i>	<i>D</i>	
	<i>A</i>	0.65	0.29	0.05	0.01	
	<i>B</i>	0.23	0.55	0.19	0.13	
	<i>C</i>	0.08	0.33	0.43	0.17	
	<i>D</i>	0.01	0.12	0.22	0.64	

**Table 29:** Scheme 3: PDDS Matching, unadjusted transitions (14 states to 11 lumps)

<i>Frequencies</i>											
	<i>A</i>	<i>B</i>	<i>C</i>	<i>D</i>	<i>E</i>	<i>F</i>	<i>G</i>	<i>H</i>	<i>I</i>	<i>J</i>	<i>K</i>
<i>A</i>	237	119	46	55	15	5	3	2	0	0	0 482
<i>B</i>	114	205	96	102	21	7	4	1	0	0	0 550
<i>C</i>	47	103	154	133	47	13	10	1	1	1	0 510
<i>D</i>	54	109	132	343	131	73	22	9	0	1	0 874
<i>E</i>	19	23	46	139	114	55	38	14	1	2	0 451
<i>F</i>	6	10	13	52	67	93	46	21	4	4	0 316
<i>H</i>	4	1	4	30	25	44	60	37	3	5	0 213
<i>I</i>	1	0	4	10	7	20	35	55	10	11	0 153
<i>J</i>	0	0	2	1	0	1	4	10	10	4	0 32
<i>K</i>	0	0	1	2	1	0	4	5	3	26	1 43
	0	0	0	0	0	0	0	0	0	1	0 1
<i>Probabilities</i>											
	<i>A</i>	<i>B</i>	<i>C</i>	<i>D</i>	<i>E</i>	<i>F</i>	<i>G</i>	<i>H</i>	<i>I</i>	<i>J</i>	<i>K</i>
<i>A</i>	0.49	0.25	0.10	0.11	0.03	0.01	0.01	0.004	0	0	0
<i>B</i>	0.21	0.37	0.17	0.19	0.04	0.01	0.01	0.002	0	0	0
<i>C</i>	0.09	0.20	0.30	0.26	0.9	0.03	0.02	0.002	0.002	0.002	0
<i>D</i>	0.06	0.12	0.15	0.39	0.15	0.08	0.03	0.01	0	0.001	0
<i>E</i>	0.04	0.05	0.10	0.31	0.25	0.12	0.08	0.03	0.002	0.004	0
<i>F</i>	0.02	0.03	0.04	0.16	0.21	0.29	0.15	0.07	0.01	0.01	0
<i>H</i>	0.02	0.005	0.02	0.14	0.12	0.21	0.28	0.17	0.01	0.02	0
<i>I</i>	0.01	0	0.03	0.07	0.05	0.13	0.23	0.36	0.07	0.07	0
<i>J</i>	0	0	0.06	0.03	0	0.03	0.13	0.31	0.31	0.13	0
<i>K</i>	0	0	0.02	0.05	0.02	0	0.09	0.12	0.07	0.60	0.02
	0	0	0	0	0	0	0	0	0	1.0	0

**Figure 23** : State transition diagram for lumped chain according to Scheme 3 (PDDS matching)



- $p_{i,j} \geq 0.10$
- $0.10 < p_{i,j} < 0.25$
- $0.05 < p_{i,j} \leq 0.10$
- $p_{i,j} \leq 0.05$ ; values not shown

Arrows/connecting lines above the states represent transition to higher states; those below represent transitions to lower states or remaining in the same state. Probabilities approximately 0 are not depicted.

### 3. Secondary Analyses

Proportional odds models with random effects models were fit for all 4 situations: overall (unlumped) and for each of the 3 lumping schemes. In all cases, parameter estimates were small. Current EDSS, baseline EDSS, and age of diagnosis were statistically associated with transitions in all models (Table 30); only for the overall, unlumped model was race associated with transitions. Race, number of relapses in the past 12 months and gender were not associated with score transitions in any model (all  $p > 0.05$ ; values not reported).

For each model, the parameters consistently had negative parameter estimates for each predictor; therefore, there is a decreased probability of larger scores (decreased mobility). In general, for the enrolment and current EDSS, this means a patient is more likely to remain close to their starting or current score than move up to the higher end of the scale (higher scores). These results are consistent between each lumping scheme.

**Table 30** : Summary of proportional odds models with random effects

<b>Lumping Scheme</b>	<b>Predictor</b>	<b>Estimate (SE<sub>1</sub>)</b>	<b>p-value<sub>2</sub></b>
<i>None</i>	Current EDSS	-0.61 (0.09)	<b>&lt;0.0001</b>
	Enrollment EDSS	-1.18 (0.10)	<b>&lt;0.0001</b>
	Gender (Female)	0.11 (0.14)	0.4115
	Age of Diagnosis	-0.04 (0.01)	<b>&lt;0.0001</b>
	Number of Relapses	0.05 (0.08)	0.4689
	Race (African American)	-0.28 (0.23)	0.3495
	Race (Other)	-0.23 (0.26)	
<i>Scheme 1</i>	Current EDSS	-0.65 (0.09)	<b>&lt;0.0001</b>
	Enrollment EDSS	-1.16 (0.10)	<b>&lt;0.0001</b>
	Gender (Female)	0.08 (0.14)	0.5376
	Age of Diagnosis	-0.04 (0.01)	<b>&lt;0.0001</b>
	Number of Relapses	0.06 (0.08)	0.4676
	Race (African American)	-0.25 (0.23)	0.4133
	Race (Other)	-0.23 (0.28)	
<i>Scheme 2</i>	Current EDSS	-0.41 (0.06)	<b>&lt;0.0001</b>
	Enrollment EDSS	-1.25 (0.09)	<b>&lt;0.0001</b>
	Gender (Female)	0.07 (0.15)	0.6090
	Age of Diagnosis	-0.04 (0.01)	<b>&lt;0.0001</b>
	Number of Relapses	0.04 (0.08)	0.6585
	Race (African American)	-0.33 (0.25)	0.2896
	Race (Other)	-0.28 (0.29)	
<i>Scheme 3</i>	Current EDSS	-0.60 (0.09)	<b>&lt;0.0001</b>
	Enrollment EDSS	-1.17 (0.09)	<b>&lt;0.0001</b>
	Gender (Female)	0.12 (0.14)	0.3700
	Age of Diagnosis	-0.04 (0.01)	<b>&lt;0.0001</b>
	Number of Relapses	0.06 (0.07)	0.4524
	Race (African American)	-0.27 (0.22)	0.3389
	Race (Other)	-0.24 (0.26)	

<sup>1</sup>Standard error<sup>2</sup>Type III Test of Fixed Effects<sup>3</sup>Number of Relapses in the last 12 months leading up to baseline



## D. Discussion and Conclusion

This chapter explored the application of the Test of Lumpability and our novel, proposed Goodness of Fit test on clinician-evaluated disability scores (the EDSS) obtained from a clinical trial. We demonstrated the challenge involved with identifying a simultaneous parsimonious and useful lumping scheme that also retains the Markov property. In doing so, we further demonstrated that even with retaining this property, lumping by a specific scheme may not be the best fit to the data.

**Primary results** In the primary analyses, all transition matrices share a common characteristic of transitions being concentrated on the diagonal, confirming the inherent dependency on previous scores and that knowledge of previous disability status is highly informative for considering future disability.

In the matrix representing the overall sample, the distribution of transitions along the diagonal are concentrated between 0 and 3.5 for the current EDSS and between 0 and 4.5 for the next EDSS scores. This is an artifact of the disability inclusion criterion for the study. This pattern is also true for Scheme 1 and so some degree, Scheme 3; most likely because these states are more granular compared to Scheme 2, which produces substantially fewer states in the lumped chain. In scheme 2, we observe a more even distribution of scores along the diagonal.

The fact that the transition matrices share a common characteristic of transitions being concentrated on the diagonal, supports the inherent dependency on previous scores and that knowledge of previous disability status is highly informative for

considering future disability. Because the disease is a slowly progressing one, it is reasonable that scores do not increase drastically during the relatively short observation period; but rather, just transition back and forth between scores.

Lumps utilized were proposed with scientific justification; upon aggregation, we observed the unlumped chain was lumpable according to only one scheme (PDDS matching). However, this  $p$ -value is only nominally different from 0.05 and so the result is questionable; additionally it failed both tests of GOF. Since we have observed that the Test of Lumpability appears to be driven by sample size, the other lumping schemes might still be useful, and perhaps investigation other strategies for lumping might yield more promising results and lead to greater insights into disability progression. Because only one scheme passed the Test of Lumpability, additional work to identify other scientifically supported and clinically meaningful schemes. The state diagram is also more evidence for a useful lumping scheme, as its complexity renders it challenging to reference.

Because of the study design, our sample has less variability, is healthier and more homogenous with lower EDSS scores. Perhaps fewer transitions in the matrix later in the scale also drove our observed results. We anticipated a problem with sparseness the upper range of scores due to the CombiRx study design; inclusion criteria required that enrolling patients not have an EDSS greater than 5.5.

Therefore, sparseness was observed for the latter parts of the scale (greater than 5.5); at the early follow-up points of the study, we observed the matrices were weighted by those scores 5.5 and less. As a result, we had to truncate our original,

unlumped chain from 20 states to 14 states. By extension, the results of the matrix are more informative for the earlier part of the disease course (lower scores); we expected fewer participants in the latter part of the scale even towards the end of the study period, because patients were recruited earlier in the disease course. Thus, little knowledge was be gained regarding the later part of the scale (higher scores).

**Secondary results** The secondary results were for covariate-adjusted transition probabilities. Consistent results regarding predictors is promising, and lends more confidence in their use as predictors of EDSS scores. However, other predictors beyond those considered be yet be useful. Investigation of the relationship between relapse and the transitions was one of our goals; however, number of relapses was not statistically associated with these transitions in the model. The relationship may yet be there, but due to the fact that all patients in the sample were on a disease modifying therapy, their number of relapses were reduced during the study period, therefore influencing the change in EDSS scores during this time.

**Future work** It was of interest to evaluate the proportional odds assumption, particularly for the Current EDSS predictor. However, due model complexity and the multitude of additional steps to resolve, this is an additional avenue of future work. While we successfully modeled transitions using our chosen predictors, we did not perform tests on adjusted probability matrices. This is a logical step to extend these analyses and will require different transition matrices for every subject at every time point in order to obtain likelihood estimates, and ultimately

perform the test of lumpability to determine if controlling for certain patient characteristics affects the lumpability of the Markov process. The large number of parameters that must be estimated (due to length of the chain) are further indication that a lumped process is desirable, as it would ease computation.

Although the starting dataset was a large sample size, it was restricted due to missing responses, a challenge with any kind of data collection. These results are not generalizable to patients with increased disability, by design of the study; this might also be true for MS patients not receiving any kind of treatment.

The primary results were for unadjusted matrices, and consideration of other lumping schemes are of interest are to be explored. It is of interest to extend these results through analyses via consideration of higher order models; consideration of other covariates; estimation of adjusted transition probabilities examination; evaluating lumpability for adjusted probability matrices; and partial proportional odds models.

## ***VI. Summary, conclusion, next steps***

It is common practice to group levels of categorical variables for statistical or sample size considerations. However, in the case of Markov chains, this aggregation can affect the properties of the resulting chain, including whether or not the new chain retains the eponymous Markov property. Because multiple lumping schemes might be appropriate for a set of data and because multiple might still retain Markov dependency, we proposed a GOF test to compare lumped and unlumped chains, and a comparator statistic to compare chains which further pass the GOF. Upon establishing the GOF tests to evaluate lumpable matrices and examining their performance in simulated, the natural extension was to explore their performance in real data. In order to observe how the methodology performs in a larger sample over a longer period time, we employed these Markov methods in NARCOMS data on the PDDS. Because the EDSS is the gold standard in MS clinical trials, we explored Markov methodology in the CombiRx data. We were interested in whether these data sets produced similar results regarding lumping schemes, and later, for predictive variables. These datasets, while created under the requirement of complete-response data (for disability scores), still had differing patterns when it came to having zero cells in their transition matrices, allowing us also to consider performance based on how the matrix was filled.

#### A. Implementation of the Test of Lumpability and Novel test development and exploration

A nuanced and important difference is between statistical appropriateness of a model and what best fits a particular data set. This understanding motivated the investigation of a GOF measure. Simulation methods demonstrated the performance of our proposed test under pre-specified settings and set the stage for investigating their applications in real data. Our chains were lumpable by design, but the results did vary according to sample size. Smaller sample sizes influenced the results; for those simulated matrices whose overall sample size was  $N=50$ , fewer were consistently lumpable ( $p < 0.05$ ). Use of smaller sample sizes were observed with smaller values for the LRTs. Bickelbach and Bode (2001) point out that there challenge between having enough transitions to obtain stable transition probability estimates, while the increasing sample size can lead to losing the Markov dependency. Therefore, further work is warranted examining the performance of the test with both smaller and increasingly larger sample sizes, and likely shall require a restriction on the sample size appropriate for whatever test is implemented.

The matrix dimension also appears to influence results; the number of cells in a matrix can have wildly varying entries, further complicating the prospect of a single score to account for so much variability. Therefore, further investigation should balance total number of transitions per matrix, sparseness, location of sparseness and number of cells (dimension). This also suggests a potential approach of moving

forward is to consider breaking up the chain by some technique that is appropriate to the process being modeled and some test of the chain is desired. The issues of balancing sample size, sparseness and, matrix dimension demonstrate a multi-layered problem.

## B. Comparison of NARCOMS and CombiRx results

*Comparison of study design characteristics* For the PDDS outcome (NARCOMS), we hoped to capture a larger cohort (compared to CombiRx) that would inform the probability of transitioning between disability states towards the later end of the scale, and so gain information regarding the nature of disability with disease progression. NARCOMS sample provides are more population general and includes patients of greater disease severity and longer disease duration, allowing us to consider stratification based on baseline enrollment. Information regarding treatment is less certain, as this is self-report. For the EDSS outcome (CombiRx), we expected issues with sparseness in the later parts of the score due to study design and inclusion criteria. We did observe this to be the case, and had to truncate our starting chain accordingly from 20 states to 14 states. Relatedly, there was no baseline disability stratification due to the nature of the study's eligibility criteria. As anticipated, we did not observe many transitions in the later part of the scale.

*Comparison of sample results and future work* Methodology was consistent between the NARCOMS and CombiRx datasets in order to draw accurate comparisons

between performance of methodology and similarity (or differences) between scales. In general, we did observe highly similar results between these 2 outcomes.

A substantial difference is the disease duration at the start of the observation period; on average CombiRx patients had the disease for less time (1.1 years (SD=3.0)) than NARCOMS patients (15.8 years (SD=8.9)). They also have a shorter observation period than NARCOMS (3 years in CombiRx, 5 years for NARCOMS). CombiRx patients are certainly are on some kind of therapy, per the protocol of the clinical trial, whereas the treatment regimen of the registry participants is going to be more variable. CombiRx patients are healthier, with less disability overall (mean EDSS 1.9 (SD=1.2)) compared to NARCOMS patients (mean PDDS 2.9 (SD=2.1)). This is in-keeping with what we expected regarding patient characteristics and also generalizability. This was further demonstrated in the concentration of the transitions within each transition matrix.

Let us consider transition activity to include only transition probabilities larger than 0.10. The CombiRx patients have similar disability status to the patients in the MI subset of NARCOMS. They also have more transition activity (that is, transitions paths to other states) than the NARCOMS HI subset (particularly for transitions from higher states to lower states), as examination of the state diagrams demonstrates (Figures 19, 22). Then we can think of this comparison as between patients with HI and MI and this difference is perhaps driven by disease duration differences: patients in the earlier stages of the disease have a less severe disease state and tend to fluctuate in their scores due to the relapsing-remitting nature of the disease.



Conversely, the NARCOMS HI patients have had the disease for longer, hence why we observe scores in the later end of the scale (while again noting that we did not restrict patients' inclusion to the NARCOMS sample based on disease status). This observation further suggests examination along these lines is warranted along.

While it is true that there is not one-to-one correspondence between disability scales, their similarity does allow for some general comparisons to be made in context of Markov chains regarding their results. The results support correspondence between the two scales. Similar associations between predictive variables and the movement between scores, particularly in how transitions are concentrated along the diagonals of all matrices considered. The state diagrams were presented for the scheme from each dataset which passed the Test of Lumpability. Examination of both demonstrate the complexity present in the models, even after lumping, and further support the need for identification of a useful lumping scheme, as such a complex figure has limited usefulness.

MS is a progressive disease whose periods of relapse and remission are its hallmark. Therefore, investigating model order would add another level of understanding to disease progression, as well as model complexity. To this point, only transition matrices at a single point in time have been considered. Stationarity should also be evaluated. Such a determination can have implications for the MS data, as it could be that a specific transition matrix is only appropriate and applicable to a certain stage or groups of stages (related to disability status) in the disease or specific disease durations. Because of the sample size and the long-term nature of the registry, there

is also great potential to tease out where and how stationarity of the Markov chain might change—that is, where it might be appropriate to have a separate Markov chain. This could be based on a combination of disease duration and disability. Such implementation may prove more difficult, however, for the CombiRx data, as the observation period is relatively short, even including the extension phase of the study. Additionally, because it is implicit in the Test of Lumpability that the original chain and the lumped chain must share the same order (first order), there is the possibility that the lumped chain is larger than order 1.

In both datasets, we observed most chains were not lumpable according to given schemes at the conventional 0.05 level. However, in both datasets, we were analyzing a large sample size, therefore it was not entirely unanticipated that we failed the Test. Larger sample sizes suggest less of a need to lump because a lot of detail is available from the dataset at a more granular level; however, as demonstrated, more states in a chain means substantially more model complexity, which can have implications for interpretability, computation, and model usefulness. This does allow for several avenues of future work. First, because we observed that the Test of Lumpability appears to be driven by sample size we can be stricture about significance level; therefore, development of guidelines on how to adjust the  $\alpha$ -level according to sample size would be useful; Second and by extension, perhaps the Test of Lumpability is not adequately penalizing for the number of parameters, because we are less likely to choose the lumped chain with an increasing sample size. Third, because we were working with large sample sizes for both the simulated and real-world data, one approach to identifying a lumpable scheme might be to

perform random sampling with replacement from the larger sample, implement the scheme, perform the Test of Lumpability and then determine what proportion of “passes” were achieved. This would, however, introduce questions about what a reasonable sample size might be and how many random samples to draw.

Importantly, a large sample is a benefit, because it does allow for accuracy model complexity regarding design and use of covariates when obtaining adjusted transition probabilities. In the context of building covariate-adjusted probability models (whether partial proportional odds or proportional odds), future work may see utilization of AIC and BIC, and potential implementation of the Vuong test, which allows comparison of non-nested models, as it has not yet been established if lumped chains are nested in their associated unlumped chains (Vuong, 1989).

It was of interest to evaluate the proportional odds assumption, particularly for the Current PDDS and EDSS predictors. However, due model and computational complexity and the multitude of additional steps to resolve, this is an additional avenue of future work. While we successfully modeled transitions using our chosen predictors, we did not perform tests on adjusted probability matrices. This is a logical step to extend these analyses and will require different transition matrices for every subject at every time point in order to obtain likelihood estimates, and ultimately perform the test of lumpability to determine if controlling for certain patient characteristics affects the lumpability of the Markov process. The large number of parameters (transition probabilities) that then must be estimated (due to

length of the chain) are further indication that a lumped process is desirable, as it would ease computation.

The structure of the matrices, themselves can influence results, including the location of zero-cells within the matrix. While lumping may get rid of zero-cells, in the context of MS their presence is informative as it is a characteristic of the disease progression. However, it could be that stratification alleviates or even eliminates this issue. Thus, it is of interest to evaluate the performance of the test with and without certain levels of sparseness. Additional considerations include imposing restrictions on the number of allowable “zero” cells, as is done in the test for stationarity (Bickenbach & Bode, 2001) and expected cell count limitations.

### C. Conclusion

Markov processes are easier to implement due to increased computing power, but still have areas of complexity that require balancing sample size (number of transitions) and matrix dimension. Lumping states can ease interpretability and computation of complex models. However, it must be justified by scientific and statistical considerations, including confirmation that chain is still Markov. While a simpler matrix could be preferable, a more appropriate and larger one might provide greater insight and superior fit to the data. Though we are in the early stages of determining useful lumping schemes for MS, there is potential to expand

this knowledge with positive implications modeling a complex outcome for a better understanding of disease progression.

## VII. Index of notation

Term	Description
$Y(t)$	Stochastic process; in this case, a Markov chain Markov chain whose probabilities are produced via logistic regression
$Y(t_i)$	Random variable in the stochastic process
$X_n = i$	Probability of being in state $i$ at time $n$
$T$ , where $T = t_1, t_2, t_3, \dots t_n$	Index set, usually time
$a_k$	Initial probability distribution; always in vector form
$S$ , where $S = \{1, 2, \dots s\}$	State space
$i$ , where $i = 0, 1, 2, \dots, s$	Current state, index
$j$ , where $j = 0, 1, 2, \dots, s$	Next state, index
$p_{i,j}$	Probability of transitioning from state $i$ to state $j$
$P^T$	Transition matrix at time $T$
$Y$	Probability estimated from logistic regression
$N$	Sample size
$n$	Referring to number of states in the observed (unlumped) matrix; used to indicate dimension and perform calculations
$m$	Referring to the number of lumps in the lumped matrix; used to indicate dimension and perform calculations

## VIII. References

1. Agresti, A. (2007). *An introduction to categorical data analysis*. Hoboken, New Jersey: John Wiley & Sons, Inc.
2. Albert, P. S. (1994). A Markov model for sequences of ordinal data from a relapsing-remitting disease. *Biometrics*, 50(1), 51-60. doi:10.2307/2533196
3. Altman, R. M., & Petkau, A. J. (2005). Application of hidden Markov models to multiple sclerosis lesion count data. *Statistics in Medicine*, 24(15), 2335-2344. doi:10.1002/sim.2108
4. Amato, M. P., & Ponziani, G. (1999). Quantification of impairment in MS: discussion of the scales in use. *Mult Scler*, 5(4), 216-219. doi:10.1177/135245859900500404
5. Anderson, T. W., & Goodman, L. A. (1957). Statistical Inference about Markov Chains. *Annals of Mathematical Statistics*, 28(1), 89-110. doi:10.1214/aoms/1177707039
6. Anton, H. (2010). *Elementary Linear Algebra* (10 ed.): Wiley.
7. Baldassari, L. E., Salter, A. R., Longbrake, E. E., Cross, A. H., & Naismith, R. T. (2017). Streamlined EDSS for use in multiple sclerosis clinical practice: Development and cross-sectional comparison to EDSS. *Mult Scler*, 1352458517721357. doi:10.1177/1352458517721357
8. Baran, R. H. (2001). *Testing the Lumpability and Conditional Independence in Markovian Models*. (PhD). American University, Washington, D.C. (3003099)

9. Bargiela, D., Bianchi, M. T., Westover, M. B., Chibnik, L. B., Healy, B. C., De Jager, P. L., & Xia, Z. (2017). Selection of first-line therapy in multiple sclerosis using risk-benefit decision analysis. *Neurology*, 88(7), 677-684. doi:10.1212/WNL.0000000000003612
  
10. Barr, D. R., & Thomas, M. U. (1977). Technical Note—An Eigenvector Condition for Markov Chain Lumpability. *Operations Research*, 25(6), 1028-1031. doi:10.1287/opre.25.6.1028
  
11. Bartlett, M. S. (1950). Recurrence times. *Nature*, 165(4201), 727-728. Retrieved from <https://www.ncbi.nlm.nih.gov/pubmed/15416811>
  
12. Beck, J. R., & Pauker, S. G. (1983). The Markov process in medical prognosis. *Med Decis Making*, 3(4), 419-458. doi:10.1177/0272989X8300300403
  
13. Bergamaschi, R., & Montomoli, C. (2016). Modeling the course and outcomes of MS is statistical twaddle--No. *Mult Scler*, 22(2), 142-144. doi:10.1177/1352458515620298
  
14. Bhat, U. N., & Miller, G. K. (2002). *Elements of Applied Stochastic Processes* (3 ed.). Hoboken, New Jersey: John Wiley & Sons, Inc.
  
15. Bickenbach, F., & Bode, E. (2001). *Markov or not Markov - This should be a question*, Kiel Working Paper No. 1086. Paper presented at the 42nd Congress of the European Regional Science Association, Dortmund, Germany. <http://hdl.handle.net/10419/2673>
  
16. Billingsley, P. (1961). Statistical methods in markov chains. *Annals of Mathematical Statistics*, 32(1), 12-40. doi:10.1214/aoms/1177705136
  
17. Cofield, S. S., Fox, R. J., Tyry, T., Salter, A. R., & Campagnolo, D. (2016). Disability Progression After Switching from Natalizumab to Fingolimod or Interferon Beta/Glatiramer Acetate Therapies: A NARCOMS Analysis. *Int J MS Care*, 18(5), 230-238. doi:10.7224/1537-2073.2014-113



18. Cofield, S. S., Thomas, N., Tyry, T., Fox, R. J., & Salter, A. (2017). Shared Decision Making and Autonomy Among US Participants with Multiple Sclerosis in the NARCOMS Registry. *Int J MS Care*, 19(6), 303-312. doi:10.7224/1537-2073.2016-091
  
19. Cohen, R. A., Kessler, H. R., & Fischer, M. (1993). The extended disability status scale (EDSS) as a predictor of impairments of functional activities of daily living in multiple sclerosis. *Journal of the Neurological Sciences*, 115(2), 132-135. doi:10.1016/0022-510X(93)90215-K
  
20. Coyle, P. K., Khatri, B., Edwards, K. R., Meca-Lallana, J. E., Cavalier, S., Ruffi, P., . . . Teri, P. R. O. T. G. (2017). Patient-reported outcomes in relapsing forms of MS: Real-world, global treatment experience with teriflunomide from the Teri-PRO study. *Mult Scler Relat Disord*, 17, 107-115. doi:10.1016/j.msard.2017.07.006
  
21. Engler, D., Chitnis, T., & Healy, B. (2017). Joint assessment of dependent discrete disease state processes. *Stat Methods Med Res*, 26(3), 1182-1198. doi:10.1177/0962280215569899
  
22. Fitzgerald, K. C., Tyry, T., Salter, A., Cofield, S. S., Cutter, G., Fox, R., & Marrie, R. A. (2018). Diet quality is associated with disability and symptom severity in multiple sclerosis. *Neurology*, 90(1), e1-e11. doi:10.1212/WNL.0000000000004768
  
23. Gauthier, S. A., Mandel, M., Guttmann, C. R., Glanz, B. I., Khoury, S. J., Betensky, R. A., & Weiner, H. L. (2007). Predicting short-term disability in multiple sclerosis. *Neurology*, 68(24), 2059-2065. doi:10.1212/01.wnl.0000264890.97479.b1
  
24. Goodin, D. (2014). The epidemiology of multiple sclerosis: insights to disease pathogenesis. In *Handbook of Clinical Neurology* (Vol. 122, pp. 231-266).
  
25. Goodin, D. S., Reder, A. T., Bermel, R. A., Cutter, G. R., Fox, R. J., John, G. R., . . . Waubant, E. (2016). Relapses in multiple sclerosis: Relationship to disability. *Mult Scler Relat Disord*, 6, 10-20. doi:10.1016/j.msard.2015.09.002

26. Healy, B. C., & Engler, D. (2009). Modeling disease-state transition heterogeneity through Bayesian variable selection. *Stat Med*, 28(9), 1353-1368. doi:10.1002/sim.3545
  
27. Hohol, M. J., Orav, E. J., & Weiner, H. L. (1995). Disease Steps in multiple sclerosis: A simple approach to evaluate disease progression. *Neurology*, 45(2), 251-255. Retrieved from <https://www.ncbi.nlm.nih.gov/pubmed/7854521>
  
28. Hohol, M. J., Orav, E. J., & Weiner, H. L. (1999). Disease steps in multiple sclerosis: a longitudinal study comparing disease steps and EDSS to evaluate disease progression. *Multiple Sclerosis and Related Disorders*, 5(5), 349-354. doi:10.1177/135245859900500508
  
29. Hutchinson, J., & Hutchinson, M. (1995). The functional limitations profile may be a valid, reliable and sensitive measure of disability in multiple sclerosis. *Journal of Neurology*, 242(10), 650-657. doi:10.1007/bf00866915
  
30. Hutchinson, M. (2016). Modeling the course and outcomes of MS is statistical twaddle--Commentary. *Mult Scler*, 22(2), 144-145. doi:10.1177/1352458516628332
  
31. Jernigan, R. W., & Baran, R. H. (2003). Testing lumpability in Markov chains. *Statistics & Probability Letters*, 64(1), 17-23. doi:10.1016/s0167-7152(03)00126-3
  
32. Jimoh, O. D., & Webster, P. (1996). The optimum order of a markov chain model for daily rainfall in nigeria. *Journal of Hydrology*, 185(1-4), 45-69. doi:10.1016/S0022-1694(96)03015-6
  
33. Katz, R. W. (1981). On Some Criteria for Estimating the Order of a Markov Chain. *Technometrics*, 23(3), 243-249. doi:10.2307/1267787
  
34. Kemeny, J. G., & Snell, J. L. (1960). *Finite Markov Chains*. New York: van Nostrand.

35. Koziol, J. A., Frutos, A., Sipe, J. C., Romine, J. S., & Beutler, E. (1996). A comparison of two neurologic scoring instruments for multiple sclerosis. *Journal of Neurology*, 243(3), 209-213. doi:10.1007/BF00868516
36. Kurtzke, J. F. (1955). A new scale for evaluating disability in multiple sclerosis *Neurology*, 5(8), 580-583.
37. Kurtzke, J. F. (1983). Rating neurologic impairment in multiple sclerosis: an expanded disability status scale (EDSS). *Neurology*, 33(11), 1444-1452. Retrieved from <https://www.ncbi.nlm.nih.gov/pubmed/6685237>
38. Learmonth, Y. C., Motl, R. W., Sandroff, B. M., Pula, J. H., & Cadavid, D. (2013). Validation of patient determined disease steps (PDDS) scale scores in persons with multiple sclerosis. *BMC Neurology*, 13, 37. doi:10.1186/1471-2377-13-37
39. Lindsey, J. W., Scott, T. F., Lynch, S. G., Cofield, S. S., Nelson, F., Conwit, R., . . . Group, C. I. (2012). The CombiRx trial of combined therapy with interferon and glatiramer acetate in relapsing remitting MS: Design and baseline characteristics. *Multiple Sclerosis and Related Disorders*, 1(2), 81-86. doi:10.1016/j.msard.2012.01.006
40. Liu, Y., Morgan, C., Hornung, L., Tyry, T., Salter, A. R., Agashivala, N., . . . Cutter, G. R. (2016). Relationship between symptom change, relapse activity and disability progression in multiple sclerosis. *J Neurol Sci*, 362, 121-126. doi:10.1016/j.jns.2016.01.034
41. Lublin, F. D., Cofield, S. S., Cutter, G. R., Conwit, R., Narayana, P. A., Nelson, F., . . . Investigators, C. (2013). Randomized study combining interferon and glatiramer acetate in multiple sclerosis. *Annals of Neurology*, 73(3), 327-340. doi:10.1002/ana.23863
42. Lublin, F. D., Cofield, S. S., Cutter, G. R., Gustafson, T., Krieger, S., Narayana, P. A., . . . Wolinsky, J. S. (2017). Long-term follow-up of a randomized study of combination interferon and glatiramer acetate in multiple sclerosis: Efficacy and safety results up to 7 years. *Mult Scler Relat Disord*, 18, 95-102. doi:10.1016/j.msard.2017.09.012

43. Lublin, F. D., Reingold, S. C., Cohen, J. A., Cutter, G. R., Sorensen, P. S., Thompson, A. J., . . . Polman, C. H. (2014). Defining the clinical course of multiple sclerosis: the 2013 revisions. *Neurology*, *83*(3), 278-286. doi:10.1212/WNL.0000000000000560
  
44. Mandel, M., & Betensky, R. A. (2008). Estimating time-to-event from longitudinal ordinal data using random-effects Markov models: application to multiple sclerosis progression. *Biostatistics*, *9*(4), 750-764. Retrieved from <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2536724/pdf/kxn008.pdf>
  
45. Mandel, M., Mercier, F., Eckert, B., Chin, P., & Betensky, R. A. (2013). Estimating time to disease progression comparing transition models and survival methods--an analysis of multiple sclerosis data. *Biometrics*, *69*(1), 225-234. doi:10.1111/biom.12002
  
46. Marrie, R. A., Cutter, G., Tyry, T., Vollmer, T., & Campagnolo, D. (2006). Does multiple sclerosis-associated disability differ between races? *Neurology*, *66*(8), 1235-1240. doi:10.1212/01.wnl.0000208505.81912.82
  
47. Marrie, R. A., & Goldman, M. (2007). Validity of performance scales for disability assessment in multiple sclerosis. *Multiple Sclerosis*, *13*(9), 1176-1182. doi:10.1177/1352458507078388
  
48. Meyer-Moock, S., Feng, Y. S., Maeurer, M., Dippel, F. W., & Kohlmann, T. (2014). Systematic literature review and validity evaluation of the Expanded Disability Status Scale (EDSS) and the Multiple Sclerosis Functional Composite (MSFC) in patients with multiple sclerosis. *BMC Neurology*, *14*, 58. doi:10.1186/1471-2377-14-58
  
49. Mieno, M. N., Yamaguchi, T., & Ohashi, Y. (2011). Alternative statistical methods for estimating efficacy of interferon beta-1b for multiple sclerosis clinical trials. *BMC Medical Research Methodology*, *11*(80), 80. doi:10.1186/1471-2288-11-80
  
50. Montalban, X., Hauser, S. L., Kappos, L., Arnold, D. L., Bar-Or, A., Comi, G., . . . Investigators, O. C. (2017). Ocrelizumab versus Placebo in Primary Progressive Multiple Sclerosis. *N Engl J Med*, *376*(3), 209-220. doi:10.1056/NEJMoa1606468

51. Mulero, P., Midaglia, L., & Montalban, X. (2018). Ocrelizumab: a new milestone in multiple sclerosis therapy. *Ther Adv Neurol Disord*, 11, 1756286418773025. doi:10.1177/1756286418773025
52. NARCOMS, R. f. M. S. (2017). Reasons to Participate Retrieved from <https://www.narcoms.org/copy-of-individuals-with-ms>
53. Palace, Bregenzer, T., Tremlett, H., Oger, J., Zhu, F., Boggild, M., . . . Dobson, C. (2014). UK Multiple sclerosis risk sharing scheme: a new natural history dataset and an improved Markov model. *BMJ Open*, 4, e004073. Retrieved from <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3902459/pdf/bmjopen-2013-004073.pdf>
54. Regnier, E. D., & Schechter, S. M. (2013). State-space size considerations for disease-progression models. *Statistics in Medicine*, 32(22), 3862-3880. doi:10.1002/sim.5808
55. Resnick, S. I. (2002). *Adventures in Stochastic Processes*. New York, NY: Birkhauser Boston.
56. Rizzo, M. A., Hadjimichael, O. C., Preiningerova, J., & Vollmer, T. L. (2004). Prevalence and treatment of spasticity reported by multiple sclerosis patients. *Mult Scler*, 10(5), 589-595. doi:10.1191/1352458504ms1085oa
57. Roskell, N. S., Zimovetz, E. A., Rycroft, C. E., Eckert, B. J., & Tyas, D. A. (2012). Annualized relapse rate of first-line treatments for multiple sclerosis: a meta-analysis, including indirect comparisons versus fingolimod. *Curr Med Res Opin*, 28(5), 767-780. doi:10.1185/03007995.2012.681637
58. Ross, S. M. (2003). *Introduction to probability models* (8 ed.). Burlington, MA: Academic Press.
59. Schwartz, C. E., Vollmer, T., & Lee, H. (1999). Reliability and validity of two self-report measures of impairment and disability for MS. North American Research Consortium on Multiple Sclerosis Outcomes Study Group. *Neurology*, 52(1), 63-70. doi:10.1212/WNL.52.1.63

60. Sharrack, B., Hughes, R. A., C, Soudain, S., & Dunn, G. (1999). The psychometric properties of clinical rating scales used in multiple sclerosis. *Brain*, 122(1), 141-159. doi:10.1093/brain/122.1.141
61. Sonnenberg, F. A., & Beck, J. R. (1993). Markov models in medical decision making: a practical guide. *Med Decis Making*, 13(4), 322-338. doi:10.1177/0272989X9301300409
62. Styan, G. P. H. (1973). Hadamard products and multivariate statistical analysis. *Linear Algebra and its Applications*, 6, 217-240. doi:10.1016/0024-3795(73)90023-2
63. Taylor, B. V. (2016). Modeling the course and outcomes of multiple sclerosis is statistical twaddle--Yes. *Mult Scler*, 22(2), 140-142. doi:10.1177/1352458515625809
64. Thomas, M. U., & Barr, D. R. (1977). An Approximate Test of Markov Chain Lumpability. *Journal of the American Statistical Association*, 72(357), 175-179. doi:10.1080/01621459.1977.10479934
65. Twork, S., Wiesmeth, S., Spindler, M., Wirtz, M., Schipper, S., Pohlau, D., . . . Kugler, J. (2010). Disability status and quality of life in multiple sclerosis: non-linearity of the Expanded Disability Status Scale (EDSS). *Health Qual Life Outcomes*, 8, 55. doi:10.1186/1477-7525-8-55
66. Vickrey, B. G., Hays, R. D., Harooni, R., Myers, L. W., & Ellison, G. W. (1995). A health-related quality of life measure for multiple sclerosis. *Qual Life Res*, 4(3), 187-206. Retrieved from <https://www.ncbi.nlm.nih.gov/pubmed/7613530>
67. Vuong, Q. H. (1989). Likelihood Ratio Tests for Model Selection and Non-Nested Hypotheses. *Econometrica*, 57(2), 307-333. doi:10.2307/1912557
68. Wang, G., Cutter, G. R., Cofield, S. S., Lublin, F. D., Wolinsky, J. S., Gustafson, T., . . . Salter, A. R. (2017). Baseline EDSS proportions in MS clinical trials affect the overall outcome and power: A cautionary note. *Multiple Sclerosis Journal*, 23(7), 982-987. doi:10.1177/1352458516670733

69. Wang, Y. C., Meyerson, L., Tang, Y. Q., & Qian, N. (2009). Statistical methods for the analysis of relapse data in MS clinical trials. *Journal of the Neurological Sciences*, 285(1-2), 206-211. doi:10.1016/j.jns.2009.07.017
70. Weinshenker, B. G., Bass, B., Rice, G. P., Noseworthy, J., Carriere, W., Baskerville, J., & Ebers, G. C. (1989). The natural history of multiple sclerosis: a geographically based study. I. Clinical course and disability. *Brain*, 112 ( Pt 1), 133-146. Retrieved from <https://www.ncbi.nlm.nih.gov/pubmed/2917275>
71. Willoughby, E. W., & Paty, D. W. (1988). Scales for rating impairment in multiple sclerosis: A critique. *Neurology*, 38(11), 1793-1793. doi:10.1212/wnl.38.11.1793
72. Wolfson, C., & Confavreux, C. (1987). Improvements to a Simple Markov Model of the Natural History of Multiple Sclerosis. *Neuroepidemiology*, 6(3), 101-115.
73. Wolfson, C., & Confavreux, C. (1985). A Markov model of the natural history of multiple sclerosis. *Neuroepidemiology*, 4(4), 227-239. doi:10.1159/000110234
74. Wolfson, C., & Confavreux, C. (1987). Improvements to a simple Markov model of the natural history of multiple sclerosis. I. Short-term prognosis. *Neuroepidemiology*, 6(3), 101-115. doi:10.1159/000110105

**APPENDIX**  
**IRB APPROVALS**



### APPROVAL LETTER

**TO:** Edwards, Lloyd J.

**FROM:** University of Alabama at Birmingham Institutional Review Board  
Federalwide Assurance # FWA00005960  
IORG Registration # IRB00000196 (IRB 01)  
IORG Registration # IRB00000726 (IRB 02)

**DATE:** 25-Jun-2019

**RE:** IRB-030911011  
Combination Therapy in MS Coordinating Center

---

The IRB reviewed and approved the Revision/Amendment submitted on 18-Jun-2019 for the above referenced project. The review was conducted in accordance with UAB's Assurance of Compliance approved by the Department of Health and Human Services.

**Type of Review:** Expedited  
**Expedited Categories:** 1, 6  
**Determination:** Approved  
**Approval Date:** 25-Jun-2019  
**Expiration Date:** 31-Jan-2020

The following populations are approved for inclusion in this project:

- Children – CRL 1

Please note:

- The PI for this protocol has changed from Dr. Cutter to Dr. Lloyd Edwards.
- Anastasia Mar Hartzes has been added to the protocol as study personnel.

Documents Included in Review:

- praf.190617
- othermisc(PIchangeemail).190617

## APPROVAL LETTER

**TO:** Cutter, Gary R

**FROM:** University of Alabama at Birmingham Institutional Review Board  
Federalwide Assurance # FWA00005960  
IORG Registration # IRB00000196 (IRB 01)  
IORG Registration # IRB00000726 (IRB 02)

**DATE:** 07-Feb-2019

**RE:** IRB-081224002  
Global Demyelinating Disease Registry (CMSC/NARCOMS Project)

---

The IRB reviewed and approved the Revision/Amendment submitted on 05-Feb-2019 for the above referenced project. The review was conducted in accordance with UAB's Assurance of Compliance approved by the Department of Health and Human Services.

**Type of Review:** Expedited  
**Expedited Categories:** 5  
**Determination:** Approved  
**Approval Date:** 07-Feb-2019  
**Expiration Date:** 06-Feb-2022

Although annual continuing review is not required for this project, the principal investigator is still responsible for (1) obtaining IRB approval for any modifications before implementing those changes except when necessary to eliminate apparent immediate hazards to the subject, and (2) submitting reportable problems to the IRB. Please see the IRB Guidebook for more information on these topics.

**The following populations are approved for inclusion in this project:**

- Children

**The following apply to this project related to informed consent and/or assent:**

- Waiver of Informed Consent

**Please note:** Anastasia Mar Hartzes has been added to the protocol as study personnel.

**Documents Included in Review:**

- praf.190109
- praf.190205.tracked