A Latent Class Analysis of Parkinson's Disease Symptoms

by

Graham Lucas Cummin

BS Biochemistry, University of Arkansas, 2017

Submitted to the Graduate Faculty of

the Department of Biostatistics

Graduate School of Public Health in partial fulfillment

of the requirements for the degree of

Master of Science

University of Pittsburgh

2019

UNIVERSITY OF PITTSBURGH

Graduate School of Public Health

This thesis was presented

by

Graham Lucas Cummin

It was defended on

December 2, 2019

and approved by

Thesis Advisor:
Jeanine Buchanich MEd, MPA, PhD
Research Associate Professor
Department of Biostatistics Deputy Director Center for Occupational Biostatistics and Epidemiology
Graduate School of Public Health
University of Pittsburgh

Committee Member:
Ada Youk PhD
Associate Professor
Department of Biostatistics Director MS Program Biostatistics
Graduate School of Public Health
University of Pittsburgh

Committee Member:
Lana Chahine MD
Assistant Professor
Department of Neurology Michael J Fox Foundation
School of Medicine
University of Pittsburgh

Jeanine Buchanich MEd, MPA, PhD

A Latent Class Analysis of Parkinson's Disease Symptoms

Graham Lucas Cummin, MS

University of Pittsburgh, 2019

Abstract

Parkinson's disease is a neurological disease in which the dopamine releasing brain cells degenerate die and are not replaced. Affecting mostly persons older than age 60, in the US population the fraction above 60 is nearly 40 percent. This population is also growing more elderly. The public health importance of correctly assessing Parkinson's disease and the accompanying symptom burden in order to effectively and efficiently treat the growing elderly population in the US and in order to keep costs and expectation managed is high. The ability to identify clusters of symptoms could improve awareness of how to treat and counsel patients. Latent Class Analysis is a method which can be used to predict classes, or clusters, and which can be used with categorical outcomes. In this thesis, the Parkinson's symptoms were clustered into four classes characterized in part by sex and age of the patient. Unique symptoms predicted at greater than 50% were identified for three of these classes, the first and reference class reported very few symptoms. Relative to the first class, the second class was more likely to have a younger age at onset, but was not more likely to be male or female, and uniquely reported Mood swings and depression (76%). The third class was more likely to be male, but was not more likely to be older or younger at age of onset, and uniquely reported difficulty standing from a chair (73%). The fourth class was more likely to be female and to be younger at age of onset relative to class 1, and uniquely reported 4 unique symptoms, sweating (62%), muscle spasm (70%), hot flashes or chills (57%) and persistent dull pain (59%).This LCA model predicts

divisions across gender and sex in the specific symptoms and their associations in keeping with clinical expectations.

Table of Contents

List of Tables

List of Figures

Preface

With acknowledgement to the Michael J. Fox Foundation for the data, and to the patience of my committee members.

1.0 Introduction

Parkinson's disease is a neurological disease in which certain brain cells responsible for dopamine release progressively degenerate, die and are not replaced. Parkinson's disease (PD) is classified as a motor system disorder, however non-motor symptoms are also experienced by persons with PD. It is a chronic disease, which progressively gets worse as the brain degenerates further. The majority of people with PD are affected after age 60, although diagnosis and treatment may begin much later. This affects treatment significantly, because much of the affected brain cells are already lost by the time diagnosis is made and treatment performed.

PD is classified by three main symptoms; tremor, stiffness which is also called rigidity, and slowness of movement which is also called bradykinesia. In the United States, an estimated 500,000 thousand persons have already been diagnosed with PD, and each year roughly 50,000 more are diagnosed. Parkinson's disease affects the elderly more than the young, as most diseases do. The fraction of the US population above 60 is near 40%, and the median age is about 38 years old. Both of these proportions are expected to continue to increase. And as the age of the US population increases, it is not unreasonable to expect that the number of persons with PD will also increase. From a public health perspective, this aging will increase the health burden that each state is asked to shoulder. Therefore, continuing to identify treatments and improving diagnostic techniques is vital to maintaining the ability to care for the population as the health burden increases. The purpose of this analysis is to identify possible classes of patients

and which symptoms are predicted by the classification, potentially improving awareness of how to treat and counsel patients. Specifically, the assumption that sex and age at onset play a role in symptom burden and symptom type is used to define the potential classification.

The causal chain by which Parkinson's disease occurs is unknown, although the incidence of Parkinson's Disease is partly explained by genetic predisposition. Treatment consists of artificially supplying the dopamine that the brain of a person with Parkinson's can no longer produce. The mainline treatment, Levadopa, is a dopamine precursor which enters brain cells, and is there cleaved to produce dopamine. It is an agonist derived from the amino acid tyrosine. An estimated 75% of persons with Parkinson's are prescribed Levadopa, which can be taken in concert with other drugs intended to treat Parkinson's disease. Levadopa is often combined with carbidopa, a molecule with a similar structure which acts as an enzyme inhibitor to ensure delivery of Levadopa into brain cells. Carbidopa therefore interfers with attempts to enzymatically digest Levadopa in the bloodstream and outside the brain, allowing more time for Levadopa to pass through the blood-brain barrier..

1.1 Methods

The method used in this analysis is called latent class analysis, the software is from the R package poLCA by Drew Linzer. It fuses several components of statistics in its estimation, including contingency tables, Bayes theorem, and likelihood theory.

The core motivation of latent class analysis is to make an attempt to explain the dependence of a set of variables. Dependence can often be inferred by examining a set of variables. A mathematical approach to determine whether there is a relationship or not is the

purpose of LCA. LCA may be used for categorical variables, making it similar to factor analysis. Thus, this method can be used to classify the presence or absence of symptoms.

Two other methods of analyzing the dependence of the data were assessed. Factor analysis was not suitable, as the outcomes in this data are categorical and not continuous. Another possibility was a tree model, however implicit in a tree model is an order or progression, which could not be justified with this data.

LCA posits that there is a latent or hidden variable which explains the apparent relationship of a set of variables. It does not name this variable, although often the expectation is that the particular disease state of a person, with their genetic and environmental aspects considered is that hidden variable which explains the relationship.

A latent class model, or LCM, estimates the observed joint distribution of indicator variables as a weighted sum of R cross classification tables of j by I dimensions. R is the number of classes and is fixed beforehand by both theoretical expectations and then confirmed retroactively by model fitting.

The LCA algorithm estimates the probability that certain outcomes, denoted j, are produced by some individual in a certain class, assuming conditional independence of the different Y potential outcomes within each jth object's outcome given class membership. This is represented formally as

$$f(Y_i; \pi_r) = \prod_{j=1}^{J} \prod_{k=1}^{K_j} (\pi_{jrk})^{Y_{ijk}}$$

Where $Y_{ijk} = 1$ if the k^{th} response is given to the j^{th} variable by the i^{th} respondent, and is 0 otherwise.

Π_{jrk} is the class conditional probability under which an observation in class 1,..R produces the k^{th} outcome on the j^{th} variable. This implies that $\Sigma\ \Pi_{jrk}=1$ within each class for each indicator variable, by definition of the probability space.

In this analysis, there were j=19 indicator variables, and each indicator variable was dichotomous with K=2. R was thought to be 4, theoretically based on the categories of the cross product of sex and age at onset. There were i=1874 subjects in this analysis.

The next component of the model is denoted p_r, and is the unconditional prior probability of an individual being in a class. In other words, it is the probability that an individual belongs to class 1,…,R without taking into consideration the j indicator variables. It weights the R cross classification tables. The $\Sigma_r\ p_r=1$. However, covariates can be introduced in order to modify this prior probability. In this analysis, the covariates of sex and age of onset were included in the analysis.

Taken all together, this is combined to be the pdf of the joint distributions across all classes.

$$P(Y_i|\pi,p) = \sum_{r=1}^{R} p_r \prod_{j=1}^{J} \prod_{k=1}^{K_j} (\pi_{jrk})^{Y_{ijk}}$$

Where P_r and Π_{jrk} are estimated from the data.

In the poLCA software package, the probabilities are estimated using expectation maximization. The log likelihood of the pdf of the joint distributions across all classes, given above, is maximized with respect to pr and π_{jrk}. Expectation maximization calculates the estimated values by picking an arbitrary value for pr and π_{jrk}, calculating the missing class membership probability using bayes theorem, and then updating both of the estimates with the

mean value of the respective posterior probabilities. This is repeated until the log likelihood is maximized.

There are a few restrictions to keep in mind with this method. The number of parameters can be increased by three different conditions. The number of outcomes within each indicator variable (K_j), the number of classes R and the number of indicator variables J. The increase in the number of parameters follows this formula $R\Sigma_i(K_j-1)+(R-1)$. Secondly, that the expectation maximization formula may find only a local maxima instead of a global maxima, so multiple iterations with different start points are advised in order to find the true global maximum. This can be achieved by instructing the algorithm to restart multiple times.

Once these are estimated, the posterior probability of each individual's class membership conditioned on the data is calculated using Bayes formula, here given formally as

$$\hat{P}(r_i|Y_i) = \frac{\hat{p}_r f(Y_i;\hat{\pi}_r)}{\prod_{q=1}^{R} \hat{p}_q f(Y_i;\hat{\pi}_q)}$$

Assessing model fit can be decided by theoretical considerations, but also influenced by post hoc analytical methods. Primarily Akaike and Bayesian information criteria are used, however in some cases chi square or G squared criteria can be used. All four methods are calculated automatically by the poLCA command. For this work, BIC was used to confirm theoretical considerations which were the primary influences on model development. The BIC is calculated as

$$BIC = -2\Lambda + \Phi \ln N$$

Where Λ is the maximum log likelihood of the model, Φ is the total number of parameters.

1.2 Application of Method

Data comes from the Michael J Fox foundation which were made available in two files. The foundation collected data from surveys which were filled out online. The main data used comes from the wearing off questionnaire 19, (WOQ 19) which is a series of 19 questions which ask about the presence of certain symptoms associated with Parkinson's. Data on demographics was also collected as part of the survey. From the WOC 19 questionnaire, further combinations and designations of categories were created. One of the categories is medication responsive, which was a category that indicated if a subject had answered yes to having any symptom, and then also answered that the specified symptom was alleviated by medication. The survey collected information on whether a person was formally diagnosed with Parkinson's disease and when, however the symptoms were self-reported.

The two files sent were merged by the unique patient identifier. The merged file was checked for duplicates and errors. Patients who stated they had Parkinson's disease but who reported a diagnosis condition of false were removed from the sample. Any patients who did not answer a single question from the WOC 19 questionnaire were also removed. There were also patients who reported developing Parkinson's disease at an age of less than 1, or who did not report their age of onset. These were also removed. The starting sample was therefore 2107 patients.

A new category was created for medication responders. These were patients who reported having any symptom, and then also reported that that symptom was alleviated by medication. The threshold for inclusion in this category was 1. To illustrate, if a patient reported having tremors and also 6 other symptoms, and only tremors was mitigated by medication, that person would be designated as medication responsive.

Summary statistics were then calculated for the sample. Age of onset of Parkinson's disease, sex, ethnicity, education, race, age when taking the survey and the duration of Parkinson's disease were summarized in table(1). Their means, sd and counts were described. T tests and chi square tests were performed to check whether the means of various sub groups were different. Age of onset was split into 2 categories, younger age of onset of <=55 or older age of onset, >55. Age at survey was split into two groups as well, demarcated by age <= 67 or > 67.

Latent class analysis was used to estimate latent class membership. Latent class is an unobserved variable or state which is theoretically thought to explain the dependence of the indicator variables. Initially, it was estimated using the medication responsive variable for each symptom of each patient as indicator variables. That is, each patient had a value for being medication responsive in a particular symptom or being non-medication responsive for that symptom. A third variable was created, non symptomatic. This variable was used to designate patients who did not report a symptom and did not report that symptom alleviated. An LCA was run with these categorization as well. These LCA's were run without covariates, and then with sex and age of onset as covariates both individually and together. Two, three and 4 class structures were assumed.

Next, LCA's were estimated with the sample subdivided by sex. These were run with no additional covariates. The created variables of med responsive/non-med responsive, and then med responsive/non med responsive/asymptomatic were used as the indicator variables.

Next, LCA's were estimated on all 19 symptoms only. That is, instead of using the data on medication responsiveness to predict latent class, the data on only presence or absence of a symptom was used. From 2 to 8 classes were assumed, and again sex and age on onset were used individually and together as covariates.

Three of the symptoms included in the WOC 19 are the definitive symptoms used to diagnose PD. Tremor, slow movement and stiffness. Latent class models estimated with these 3 included among the indicator variables showed a separation by presence or absence of combinations of the 3 main symptoms. As diagnosis of Parkinson's Disease relies upon these 3 symptoms, it is somewhat axiomatic that a person with PD will have these symptoms in some combination. The average persons in this data had at least 2 of these symptoms. Interest in the remaining symptoms and how they are grouped together was the focus of this analysis, thus these symptoms were dropped from the indicator variable list. Given that the latent class is assumed to be some particular disease state of Parkinson's disease, and that the disease is diagnosed by the presence of the above mentioned symptoms, removal of the symptoms from the indicator variable set is justified on the grounds that the latent class already contains the information that these variables convey, as PD is not diagnosed on the basis of any symptoms but these. Comparison of the predictions made by excluding and by including these symptoms is materiel for future analysis. New LCA's were estimated, with class number ranging from 2 to 8 with sex and age on onset as covariates, both individually and together.

Further analysis showed that there was large number of persons, who were skewing the latent class analysis. Consistently, across various numbers of assumed classes, there was a single class which had the highest predicted probabilities for all symptoms. This made separation of the remaining classes unclear. These persons had an average of 13 symptoms, compared with an average of 6 for the remainder of the population. They developed Parkinson's disease 3 years earlier than average and had the disease for 1.5 years longer. At roughly 11% of the sample, this was not merely a group of a few outliers. However, it also did not appear that they could be representative of the average Parkinson's patient. Given the partial genetic causality of

Parkinson's disease, it is possible these persons have a genetic profile which is more susceptible to developing PD. While this population should be looked at more closely, for this analysis it was decided that the focus should be on a more general population. Excluding this large group which differed markedly from the average allowed for the smaller variances within the remainder to be examined.

Therefore the percentage of subjects with more than 10, 11 and 12 symptoms were calculated. Subjects with 12 or more symptoms were dropped from the sample. This gave a sample size of 1874. This sample was again examined, both with all 19 symptoms and without the top 3 symptoms of tremor, slow movement and stiffness. LCAs were estimated, with from 2 to 8 classes and with sex and age of onset as covariates. LCAs were also estimated with the variables of med responsive/non-med responsive and with the med responsive/non-med responsive/asymptomatic categorization.

From this new sample, the summary statistics were re calculated, as well as the average numbers of symptoms and of med responsive symptoms per predicated class.

For data analysis and discussion purposes, the percentage of med responsive symptoms per sub group, and divided by sex and age of onset were calculated. This resulted in 16 subgroups, whose percentage of med responsiveness for each symptom was calculated. These calculations were reflected in a heatmap. Further heatmaps were generated for the class division by demographics..

2.0 Results

Table 1 Cohort Characteristics

Characteristics	Group 1: N=2107	Group 2: N=233	Group 3: N=1874
Mean Age (SD)	66.6 (8.5)	63.6 (9.6)	66.9 (8.3)
Sex			
Male	1102 (50.0)	104 (44.6)	998 (53.3)
Female	1101 (50.0)	129 (55.4)	876 (46.7)
Mean Age at onset (SD)	60.4 (9.7)	57.2 (10.0)	60.8 (9.3)
Race			
White	2058 (97.7)	224 (96.1)	1858.8 (99.2)
Non-white	14 (0.7)	6 (2.6)	8.1 (0.4)
Unknown	35 (1.7)	3 (1.3)	
Education			
<Bachelors	655 (31.1)	113 (48.5)	542 (28.9)
Bachelors	655 (31.1)	61 (26.2)	594 (31.7)
>Bachelors	793 (37.6)	59 (25.3)	733 (39.1)
Unknown	4 (0.2)		5 (0.3)
Disease duration median years	4.9	5.6	4.8

This table summarizes the different demographic statistics calculated for each sample that was considered in building the data from which to construct the latent class algorithm.

Table 2 Characteristics by sex of the analytical group

Characteristics	**Group 3 N=1874**	**Group 3: N=1874**
	Females (N=876)	Males (N=998)
Mean Age (SD)	66.7 (8.1)	67.2 (8.5)
Mean Age at onset (SD)	60.5 (8.8)	61.2 (9.7)
Disease duration median years	5.0	4.5
Race		
White	871.6 (99.5)	987.2 (98.9)
Non-white	3.0 (0.3)	5.1 (0.5)
Unknown		
Education		
<Bachelors	298.2 (34.0)	244.1 (24.5)
Bachelors	262.1 (29.9)	332.2 (33.3)
>Bachelors	313.2 (35.7)	420.2 (42.1)

This table summarizes the demographic statistics within the final sample that was used in constructing the latent class algorithm, split be sex.

2.1 LCA Results

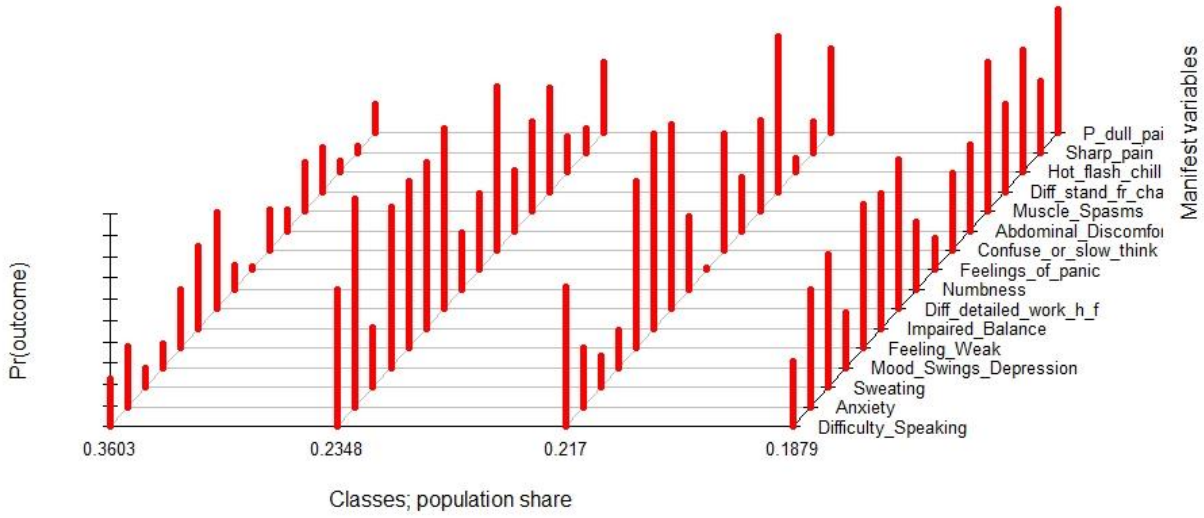

Figure 1 Graphical results of the poLCA command with 4 classes, sex and age of onset as covariates.

This is an image of the latent class analysis of the 1874 person sample with 16 indicator variables. X axis is proportion of the sample, z axis is the symptoms/indicator variables and the y axis is the probability of responding yes to the presence of the symptom.

Table 3 Symptom classes and characteristics of subjects in each class

Class demographics/defining symptoms	Class 1 N=684 (%)	Class 2 N=451 (%)	Class 3 N=406 (%)	Class 4 N=333 (%)
Male	386 (56.43)	282 (62.53)	280 (68.97)	50 (15.02)
Females	298 (43.57)	169 (37.47)	126 (31.03)	283 (84.98)
Age at onset ≤55 years	137 (20.03)	130 (28.82)	81 (19.95)	114 (34.23)
Age at onset >55 years	547 (79.97)	321 (71.18)	325 (80.05)	219 (65.77)
Sweating				62
Mood swings and/or Depression		76		
Muscle spasm				70
Difficulty standing from chair			73	
Hot flashes/chills				57
Persistent dull pain				59

Post algorithmic creation of classes with the proportion of the priors(sex and age at onset) included in the analysis, and with the symptoms unique to the class (predicted to occur in >50% only for the specified class) shown in gray-shaded cells.

Table 4 Model fit statistics

# of classes	BIC
2	34263.67
3	33963.1
4	33800.61
5	33800.1
6	33813.2

Comparison of the BIC of various class models. Fit will usually improve with an increase in the number of classes. The 4 class and 5 class BIC are very similar, the decision was made to use the 4 class model as it was consistent with clinical experience.

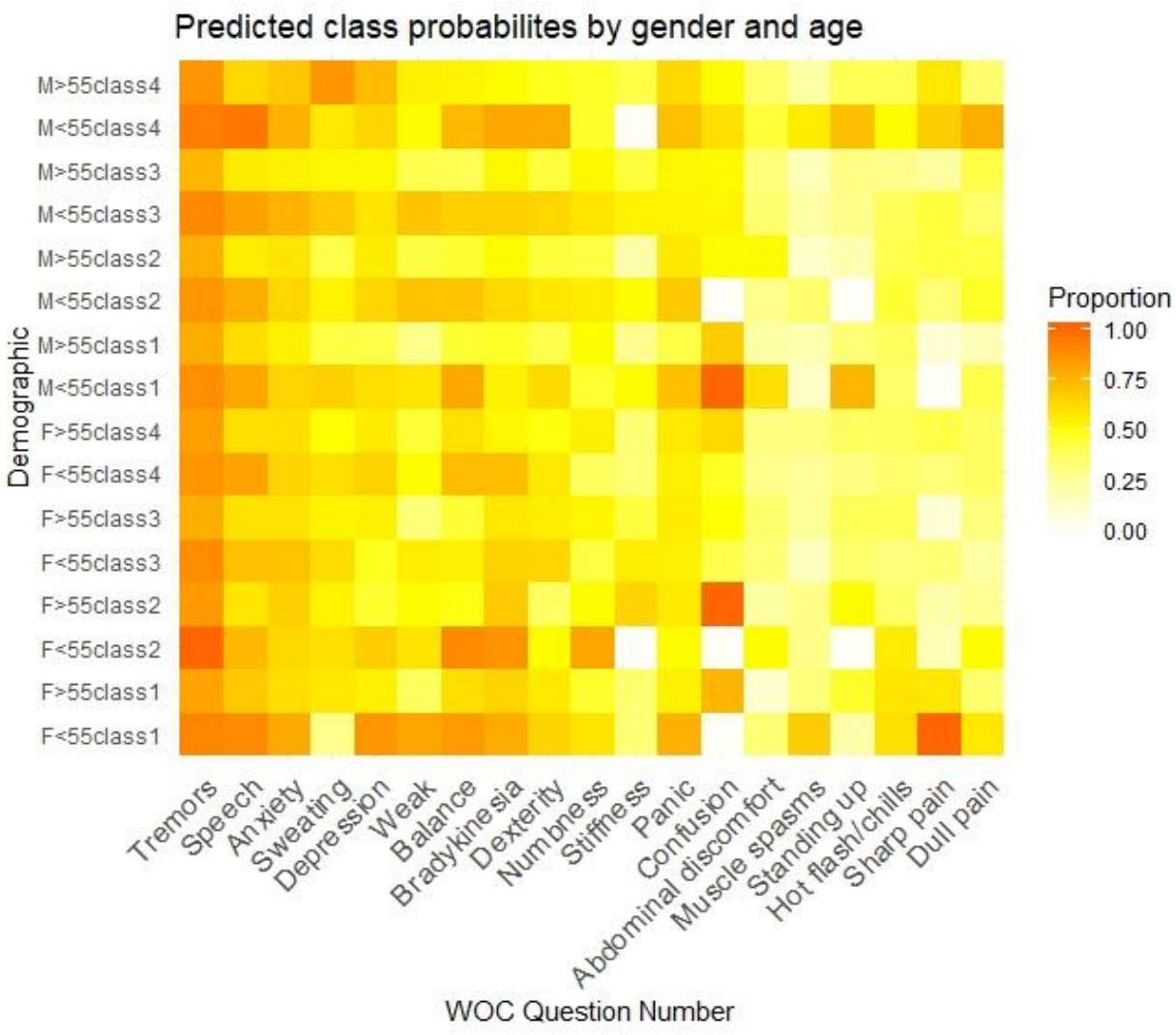

Figure 2 Heatmap of the probabilities of symptoms in each latent class by gender and age at onset

This figure is labeled in order gender, age at onset, predicted class on the Y axis. Tremor, slow movement and stiffness were included but are calculated proportion and are not predictions of probability.

Table 5 Unique symptom prevalence proportionately with class comparisons

	class1	class 2	class 3	class 4
Sweating	0.03	0.07	0.03	0.12
Mood swings/depression	0.01	0.19	0.03	0.05
Muscle spasms	0.08	0.1	0.09	0.14
Difficulty standing up	0.001	0.11	0.17	0.08
Hot flashes and chills	0.02	0.04	0.01	0.12
Persistent dull pain	0.05	0.07	0.09	0.11

Assuming the sample in this study is representative of the general population, this table shows the proportion of unique symptom presence with respect to the class size. This gives an adjusted estimate of how the presence of a symptom would correspond to latent class membership in a clinical setting.

Based on clinical expectations, a four class latent model was chosen as the best explanation of the data. Because of the high proportion of cells with no response in the 19 by 19 matrix, a χ^2 goodness of fit test could not be used, nor could the similarly restricted G^2 test. Neither was entropy an adequate measure. The nature of the LCA estimation is such that increasing the number of classes will also result in better fit according to the AIC, BIC and entropy criterion, therefore it is recommended by Dr. Linzer that practical criteria should be used primarily to evaluate the best classification. In this case, because of the use of age of onset and sex as covariates a 4 class model was expected to be the best explanation. This was confirmed using BIC as a measure of model fit.

In each class, the proportion of symptoms predicted to be greater than 50% were used to define the class. This cutoff was decided upon because of the binary nature of the symptoms, where either a person has the symptom or does not. A probability of greater than 50% indicates that the symptom is not neutral with respect to its occurrence. Probabilities less than 50%, while providing some evidence of lesser symptom likelihood, were not used as classifiers due to the

difficulty of logically showing that absence of a symptom was a defining characteristic. However, subsequent analysis combined with insight into Parkinson's progression could result in using both the absence of symptoms and their presence to correctly identify the likely symptom burden of a patient.

The model resulted in 4 patterns. The first class, which was used as the default for comparison with the covariates, was asymptomatic. Relative to the first class, the second class was more likely to have a younger age at onset, but was not more likely to be male or female. The third class was more likely to be male, but was not more likely to be older or younger at age of onset. The fourth class was more likely to be female and to be younger at age of onset relative to class 1.

3.0 Discussion

The initial questions which prompted the analysis of this data were more concerned with the responsiveness of patients to their medication. It was thought that the pattern reflected there would be illustrative and provide a better understanding of how patients who had PD were potentially subdivided, and could therefore improve physician's understanding of this disease. However, after some analysis it was discovered that the pattern of medication responsiveness was not particularly illuminating. It was also apparent, and discussed, that whatever pattern a patient would have in terms of medication responsiveness would be dependent on the symptoms the patient had reported. As both sets of data, the medication responsiveness and symptom burden were available, it was decided to focus on the underlying symptoms for pattern analysis.

Under the 4 class model, a unique pattern of symptom presence was seen which was associated with the various classes and covariates. As written above, class 1 was largely asymptomatic, with no symptom reported at greater than 50% prevalence. Class 2 uniquely reported Mood swings and depression (76%). Class 3 uniquely reported difficulty standing from a chair (73%). Class 4 uniquely reported 4 unique symptoms, sweating (62%), muscle spasm (70%), hot flashes or chills (57%) and persistent dull pain (59%). This is reported in table 3.

Although many of these symptoms were reported in all classes, the prevalence did not exceed 50%, and so they were not taken to be representative of the theoretical classes. Now, the classes were of different sizes. Assuming that the proportions are truly representative of the

population, the proportion of these symptoms adjusted for class size is given in table 5. Some of the unique symptoms do not have a much larger proportion once estimated class size is considered, however, the estimated class size is based on the data, and it is unknown exactly how representative of the actual population it is. These results should be taken as guidelines, on their own are not enough to ensure a complete identification. However, this is a start to the process of completely identifying the potential different sub classes within Parkinson's disease. These symptoms clustered together in ways which were unique and identifiable. Although the most common type of person with Parkinson's will be one with a low symptom burden, the presence of certain symptoms in combination with sex and age of onset can indicate the potential existence of other symptoms, making it easier for a physician to prepare for the future needs of the patient and to be alert for early indications of growing symptom burden and severity.

The clinical expectation that sex and age at onset would likely affect symptom burden and the interest in what the divisions look like which prompted this study has now been quantified and can be compared with future work and experiences.

Bibliography

Linzer Drew A., Lewis Jeffrey B. (2011). poLCA: An R Package for Polytomous Variable Latent Class Analysis. Journal of Statistical Software, 42(10), 1-29. URL http://www.jstatsoft.org/v42/i10/.

MacCutcheon, Allan L. (2002) Latent Class Analysis. Newbury Park, CA. Sage Publ.

Agresti, Allen (2002) Categorical Data Analysis, 3rd edition Hoboken NJ John Wiley and Sons INC

RStudio Team (2015). RStudio: Integrated Development for R. RStudio, Inc., Boston, MA URL http://www.rstudio.com/.