Fox Insight data exploration and automatic Parkinson's disease stage assesment from patient health and survey data.

Marek Travnik TRAVNIK@MIT.EDU
Department of Aeronautics and Astronautics
Massachusetts Institute of Technology
Cambridge, MA, United States

Nicholas Ramirez RAMIREZN@MIT.EDU
Department of Electrical Engineering and Computer Science
Massachusetts Institute of Technology
Cambridge, MA, United States

Gokul Kolady GOKULK@MIT.EDU
Department of Electrical Engineering and Computer Science
Massachusetts Institute of Technology
Cambridge, MA, United States

1. Introduction

This article leverages patient reported data on Parkinson's Disease Smolensky et al. (2020) from a recently published survey database by Fox Insight[1].

The initial goal of the project was to do progression modelling of Parkinson's Disease (PD), however, after discussions with our clinical mentor we decided that because the Fox Insight dataset is relatively new and unexplored, doing disease progression immediately would be premature. For this reason the focus of the project is the exploration of this dataset and exploration of methods for imputation of standardized PD progression markers.

The widely accepted measure of a patients PD severity is the UPDRS (Unified Parkinson's Disease Rating Scale) Goetz et al. (2008). However, obtaining these scores for a single patient requires an examination by a physician. The Fox Insight database provides a uniquely large sample of PD patient surveys with more than 50k patients and 300k data samples in total. Only a very small subset of these (about 200 patients), however, contain UPDRS evaluations. An automated method that could label patient survey samples with estimated UPDRS scores could provide opportunity for progression modelling of PD with estimated UPDRS scores being used as progression markers.

The main contributions we wish to provide are hence:

- Explore the Fox Insight dataset and describe it's characteristics
- Validate that the dataset provides useful information about PD progression
- Experiment with methods for imputing UPDRS scores for patient surveys.

1. https://foxden.michaeljfox.org/insight/explore/insight.jsp

2. Related Work

As eluded to earlier, the widely accepted scale for PD progression is the Unified Parkinson's Disease Rating Scale (*UPDRS*) Goetz et al. (2008). This scale is what a lot of our work in this article revolves around. The scale consists of four parts. The first part assesses a patients non-motor aspects of daily living. The second part assesses motor aspects of daily living (this is the part we mainly focus on in this article). The third part constitutes a motor examination of the patient and concludes with a commonly used assessment of the patients mobility impairment: the Hoehn and Yahr (*H&Y*) score. Part IV then assesses specific motor complications.

There is very limited literature on leveraging ML techniques on PD patient survey data. Most articles have focused their ML effort on other sources of data such as speech Frid et al. (2014), motion sensors Shetty and Rao (2016) or MRI scans Salvatore et al. (2014).

The paper Doshi-Velez et al. (2014) utilizes clustering to detail co-occurrence of medical conditions for patients with autism. This paper inspired us to use clustering as a possible method for projecting patient progression. Since our main goal changed, we decided it is still a method worth trying for imputing UPDRS scores.

3. Data Exploration

As mentioned in the Introduction, one of the main contributions this projects aims to provide is an exploration and validation of the Fox Insight data. In this section we explore the characteristics of the data in subsection 3.1, we provide a closer look at some feature categories used for later analysis in subsection 3.2 and we validate that Fox Insight data provides useful information about patient PD progression in subsection 3.3

3.1. General information and demographics

The Fox Insight database is freely downloadable and updated with additional data every month. As of May 2021 there are 51k participants in the program, totalling 300k responses.

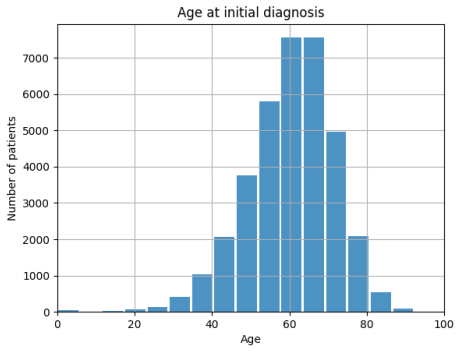

Figure 1: Distribution of patient's age at first diagnosis with PD. Mean 60.3 years.

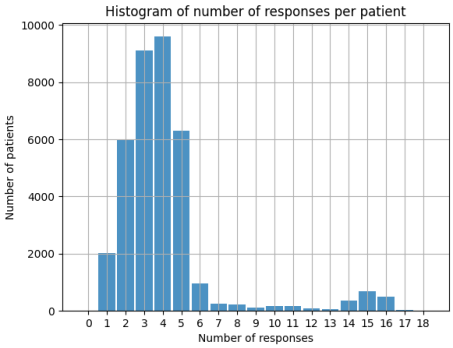

Figure 2: Distribution of number of responses per patient.

There are 36835 patients who are currently diagnosed with PD. In this paper, for most purposes only data for patients with PD diagnosis is used and the control cohort is discarded. 54.6% of patients diagnosed with PD are male, 44.1% are female and 1.3% did not specify. The plots shown here only relate to patients who were diagnosed with PD. The age distribution at the time of diagnosis with PD can be seen in Figure 1 and has a mean diagnosis age of 60.3 years. The number of responses per patient can vary significantly as shown in Figure 2, showing a distribution with mode 4 and a heavy right tail. The maximum number of points per patient is 17.

The time between visits (or time between responses) can also vary significantly as shown in Figure 3. The most common interval is 1 year. This fact then reflects in total observation periods as shown in Figure 4 which shows peaks at 0,1,2 and 3 years. An observation time of 0 means that a given patient has only completed the survey once.

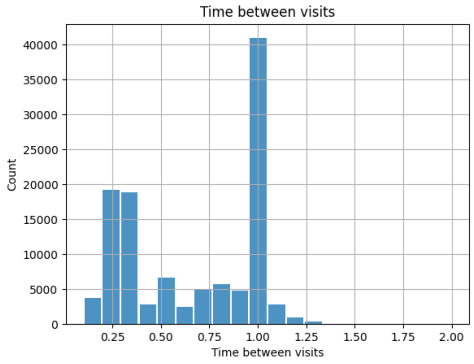

Figure 3: Dist. of times between visits.

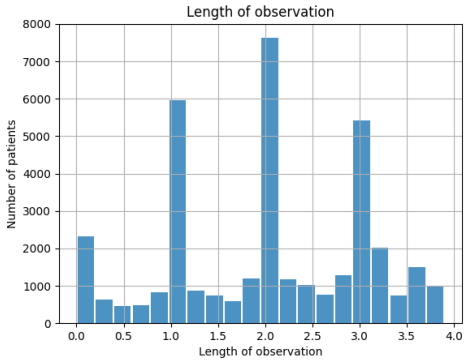

Figure 4: The length of total observation windows per patient.

3.2. Variables

Fox Insights data includes 58 categories of patient information, with a total of 5303 variables. For our effort we only use a subset of these which pertain to the tasks at hand. We divide the variables we use into three main groups.

General information				**Patient Health**			
Sub-categ.	***#Var***	***Abbr.***	***Scale***	***Sub-categ.***	***#Var***	***Abbr.***	***Scale***
General	1	*GEN*	Y/N	Current Health	86	*CH*	Y/N
About You	24	*AY*	mixed	Health History	104	*HH*	Y/N
Registration	23	*REG*	mixed	Medications	26	*MED*	Y/N
				Medications (PD)	58	*MPD*	Y/N

Table 1: General information and Patient health variable sub-categories with the number of variables, the reported scale and an abbreviation. Y/N : Yes/No

Table 1 shows the sub-categories that fall into the General Information and Patient Health groups. This table includes the number of variables in each sub-category and an abbreviation for each of them which will be used throughout the paper. Table 2 then shows the same but for sub-categories related to patients personal experience with PD. These are generally survey questions asking a patient if they have difficulty with a given activity, e.g. writing.

Personal PD experience			
Sub-category	***#Var***	***Abbr.***	**Scale**
Brief Motor Screen	10	*BMS*	Y/N
Your Cognition and Daily activities	15	*CDA*	0-5
Your Daily Living	8	*DL*	0-5
Your Movement Experiences	14	*ME*	1-5
Your Non-Movement Experiences	30	*NME*	Y/N
Physical Experiences	5	*PE*	0-5

Table 2: Survey sub-categories relating to Personal PD experience of patients.

In addition, the target variables used are UPDRS stage 2 (*UPDRS2*) mean or total scores and Hoehn and Yahr (*H&Y*) scores. Both *UPDRS2* and *H&Y* are only available for a very small cohort of patients, namely 222 individuals.

3.3. Patient progression

One of the possible ways to take advantage of the available Fox Insight data in the future is to use it for PD progression modelling. This section does not directly relate to the main tasks executed in this paper but gives important insight into the data for future use specifically for progression modelling.

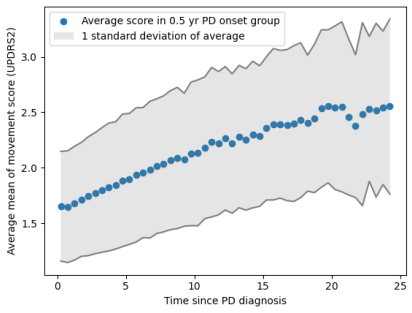

Figure 5: Progression of the mean movement score averaged over 0.5 year intervals since the diagnosis of PD.

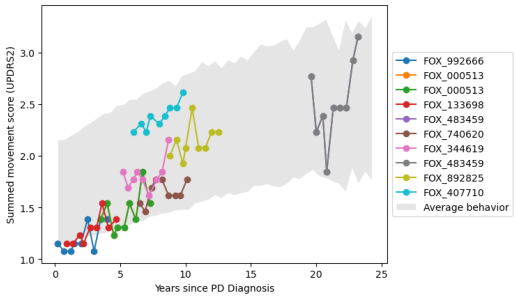

Figure 6: A sample of mean movement score progression for 10 individual patients with at least 9 visits with reported movement scores.

It will be shown in subsection 6.1 that the *ME* survey sub-category provides a good proxy for *UPDRS2* scores for a given patient. In particular, the mean score for *ME* correlates well with the total score for *UPDRS2* and is hence used as proxy. For the purpose of this exercise, a new variable called *TimeSinceDiag* was created using the age of the patient at the current visit and subtracting their age when they first were diagnosed. *TimeSinceDiag* will also be used further in modelling done in this paper. Figure 5 shows how the mean *ME* score (and by extension *UPDRS2*) tends to progress depending on the time since diagnosis. In particular, it shows the average mean *ME* score in 0.5 year intervals since diagnosis. A clear close to linear trend can be seen, validating that PD patients in the Fox Insight dataset do progress overtime. One can also see some sampled patient trajectories in Figure 6, showing that individual progressions can look quite different from each other, with some even going against the upward trend. These patient trajectories will be explored in further work beyond this paper.

4. Methods

Our effort is divided into three over-arching tasks. These tasks are, for the most part, independent. The only exception being that Task 2 relies on an important insight from Task 1, which is that the movement scores *ME* can be used as proxy for *UPDRS2*.

- **Task 1: Finding a proxy for standard scores.**
- **Task 2: Predicting proxy from limited patient data**
- **Task 3: Clustering limited patient data to find patterns**

These tasks are examined in subsection 4.1, subsection 4.2 and subsection 4.3 respectively.

4.1. Task 1: Finding a proxy for standard scores

Task 1 concentrated on finding a simple way to translate results found in terms of survey data into the widely accepted scale in the form of *UPDRS2*. Specifically this task ended up focusing mainly on predicting *UPDRS2* score means and *H&Y* scores from survey questions. This problem dealt with a very low number of datapoints. There are only 222 patients with *UPDRS2* and *H&Y* data, and only about half of these have corresponding personal experience survey data (< 200). There is also a suspected linearity between survey questions and *UPDRS2* scale scores (a patient expressing difficulty in a survey question is expected to reflect in an increased *UPDRS2* score).

The problem of predicting *UPDRS2* means turns out to be quite straight forward as very good performance can be achieved by using it's analogue survey sub-category *ME*. Hence for this problem simple linear regression does the job.

When it comes to predicting *H&Y* scores from survey data the problem is slightly more complicated, as there is no direct analogue in the survey. For these reasons, it seemed appropriate to use all Personal PD Experience survey sub-categories and regularized linear regression algorithms: Lasso and Ridge. Five-fold cross-validation was used to find the optimal regularization parameter λ for both of these. These methods also provide simple insight into which survey questions are most important in predicting H&Y scores.

4.2. Task 2: Predicting proxy from limited patient data

Task 2 focused on using limited patient information to predict a progression marker in the form of a *UPDRS2* proxy. By limited data we then mean only a patients medical and general information. However, it will be shown that adding a quick patient survey in the form of *BMO* can greatly imptove performance. It has also been mentioned that the movement experience section of the survey *ME* provides a good proxy for *UPDRS2*. If *ME* scores could be reliably predicted, and *ME* is indeed a good proxy for *UPDRS2*, then in essence simple medical data and general information about a patient could be used to accurately estimate a patient's stage in the *UPDRS2* scale. This would allow for an extremely easy and low-effort PD stage estimate, as it would only require a patient's medical record as input and would allow for patients to not have to fill out these proxy surveys. One of the methods that we chose to explore to accomplish this task was neural network regression. For the purposes of this project, we chose to focus on having the neural network predict the mean *ME* score (an average of a patient's answers on the survey between 1 and 5).

In terms of the set-up of the model itself, we settled on a neural network with one layer. We chose to keep the model to one layer as adding hidden layers did not seem to increase the performance of the network. In order to evaluate the network's performance, we utilized a 5-fold cross validation that used 80% of the usable data for training and 20% of the usable data for testing during each cross validation trial. Furthermore, for each trial, we trained the network on the training data for 10 epochs, as this was the smallest amount of epochs that achieved peak performance consistently (this also allowed for reasonable training time for the purposes of this experiment). Finally, we used mean squared error as the loss function, and Adam as the optimizer for training.

4.3. Task 3: Clustering limited patient data to find patterns

Another way we can attempt to discover patterns in the data and correlate these with standardized scores is to cluster patient samples based on patient information (health, general and *BMO* same as in Task 2), and then analyze the distribution of *H&Y* and *UPDRS2* scores within those clusters. Since the covariates used contain both categorical and numerical values, we used Factor Analysis of Mixed Data (FAMD) to create a representation of the data where distances are meaningful. Specifically, we reduce the original data to 300 dimensions. Our clustering method assumes that the use of FAMD will maintain/extract the core factors within the data. Specifically, it assumes that the information lost due to dimensional reduction is not significant. We used K-Means Clustering using Elkan Algorithm since we have a predetermined number of clusters we will create. For analysing the correlation to *H&Y* scores, we clustered the data into 6 total clusters (a cluster for each *H&Y* level) and then analyzed the distribution of *H&Y* scores within each cluster. Similarly, for *UPDRS2* scores, we clustered the data into 5 total clusters (a cluster for each level) and then analyzed the distribution within each cluster. As mentioned, we focused on *UPDRS2* due to its relevance and importance.

5. Experiment Setup and Pre-processing

The bulk of pre-processing done for Task 1 consisted of trimming the dataset down to samples containing *UPDRS2* and *H&Y*. For predicting *UPDRS2* only the *ME* sub-category was used. This regression was based on 88 datapoints that contain both *ME* and *UPDRS2* information. For predicting *H&Y* scores, all Personal PD experience sub-categories were included except *BMS* (there are no samples containing *H&Y* that contain *BMS*). After trimming down to samples containing *CDA*, *DL*, *ME*, *NME*, *PE* and *H&Y* we are left with 78 datapoints. An 80-20 split was created for training and testing, meaning that test evaluations are only based on a limited number of 16 samples. When it comes to the evaluations themselves, even though we are dealing with regression we have decided to create an appropriate accuracy metric for better interpretation. The prediction is considered accurate if it is within a threshold value t of the true value. Two values of t are then used to evaluate the models: 0.5, 0.75. The percentage of predictions within the threshold is reported as accuracy. Results are then compared to a baseline "model" which consists of predicting the mean *H&Y* value for all samples.

For both Task 2 and Task 3, the Fox Insight categories highlighted in Table 1 were used as covariates for the most part. Specifically, these categories included *AY*, *BMO*, *GEN*, Initial diagnosis age from *REG*, *CH*, *HH*, *MED*, *MPD*, and *TimeSinceDiag* described in subsection 3.3. There were a couple of steps performed for both sub-tasks in order to pre-process the data. Firstly, all patients in the dataset who hadn't been diagnosed with Parkinson's were removed. Next, we filled out missing values in a given covariate column for a given patient by drawing from earlier samples taken for that same patient. In addition, many patients had missing treatment (Do you receive treatment for x medical condition), limiting (Does x medical condition limit your daily activities), and congestive heart failure columns. To fill in these values, we assumed patients with no current heart conditions will not be receiving any heart medical treatment. Thus, if our variable was CurrHeartTrt, we assigned this variable to 0 if that patient CurrHeart variables summed to 0. Otherwise, we discarded the sample completely since we can't assume someone is receiving treatment for a heart condition. We applied this logic to all variables ending in "Trt", "Lim", and to "CurrHeartTypeCon". Finally, after all other pre-processing was complete, we dropped all samples/rows in the dataset that contained any missing values and standardized the covariate data.

Additionally, there were a couple of decisions made in covariate selection and pre-processing that were unique to Task 2. Firstly, our model was run on a set of data containing all of the covariates mentioned above (120,688 samples large), but it was also run on a set with covariates that did not include *BMS* (24,666 samples). The labels for each patient sample were created by averaging all of the responses within the *ME* category for that sample. The same notion of accuracy as for Task 1 is used here to evaluate versions of the model. In particular we evaluate the models with an accuracy threshold of $t = 0.5$. As a benchmark, we used a "model" that predicted the mean label across all training samples for every sample during test data prediction.

6. Results

6.1. Task 1: Finding a proxy for standard scores.

It can be seen in Figure 7 that there is a clear linear trend between a patients mean *ME* scores and their reported mean *UPDRS2* scores. Note that the correspondence is not one-to-one because *UPDRS2* is reported on a scale of 0-4 whereas *ME* is on a 1-5 scale. The R^2 score for this linear fit is 0.813. This was found sufficient to justify using *ME* as a proxy for *UPDRS2* for Task 2.

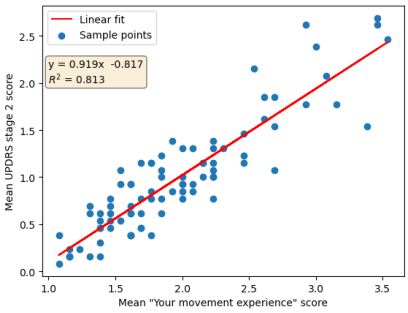

Figure 7: Linear regression showing how mean *ME* score can be used to estimate *UPDRS2* mean scores.

Method	% within 0.5	% within 0.75
Ridge $\lambda = 63.5$	63.5	75
Lasso $\lambda = 0.07$	68.8	75
Mean	50	50

Figure 8: Performance of tested methods for predicting Hoehn and Yahr scores from survey data. The column "% within 0.5" shows the percentage of predictions for a give model that fell within 0.5 of the true *H&Y* score. Similarly for "% within 0.75".

The prediction accuracies for predicting *H&Y* scores from survey data using Lasso, Ridge and regressing to the mean can be seen in Figure 8. At an accuracy threshold of 0.5, Lasso performs best with almost 70% accuracy, which is almost 20% higher than just predicting the mean value. Even though this performance is not very convincing, Lasso is still useful in providing us insight into which survey questions were most important for making it's predictions. These are reported in Table 4 in the Appendix. Note that only questions with a positive coefficient were significant enough to show, since there were only a few with negative influence with a small absolute coefficient value because most survey questions were framed such that higher answer values indicated higher difficulty due to PD.

6.2. Task 2: Predicting proxy from limited patient data

R = Regression B = Baseline	Dataset w/ *BMS*	Dataset w/o *BMS*
R: Mean CV Test Accuracy	79.34%	69.97%
R: Mean CV Per-Sample Test Squared Error	0.2023	0.2844
B: Mean CV Test Accuracy	57.16%	57.39%
B: Mean CV Per-Sample Test Squared Error	0.3942	0.4002

Table 3: Model performance for predicting *ME* scores with and without *BMS*

Table 3 shows the performance of the regression and baseline models on both the cohort of data with *BMS* in its covariates and the cohort without it. As the performance metrics show, the regression model performed significantly better than the baseline model across the board. For the cohort with *BMS*, the regression model had a 22.18% higher average test accuracy across all cross-validation trials. For the cohort without *BMS*, the regression model had a 12.58% higher average test accuracy. Furthermore, the error values for the regression model were consistently lower than those produced by the baseline. This means that the neural network did indeed pick up on meaningful trends when trained on both cohorts, and evidently did not heavily regress to the mean during prediction.

Also important to note, the network did 9.37% better in terms of test accuracy on the cohort with *BMS* than on the one without it (granted, the datasets were differently sized). This implies that the *BMS* survey acts as an important set of features in predicting the mean *ME* scores. The error was also notably lower for the *BMS* cohort.

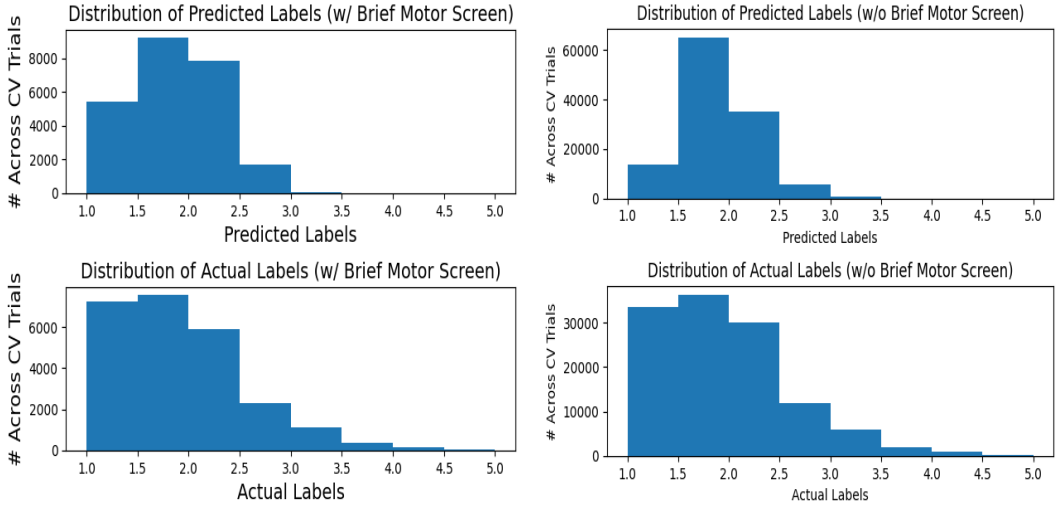

Figure 9: Predicted vs. Actual Label Distribution for dataset including *BMS* and dataset excluding *BMS*

The histograms in Figure 9 show the distributions of predicted labels and actual labels across all cross-validation test data trials for both data cohorts. For both cohorts, the relative distributions of actual labels were quite similar (the standard deviation was 0.63). The predicted label distribution for the cohort with *BMS* was more varied (standard deviation of 0.44) and slightly more spread out than the distribution for the cohort without *BMS* (standard deviation of 0.35). This closer emulation of actual label distribution by the model trained on the cohort with *BMS* would explain its superior performance. In essence, it seems as though the model trained on data with *BMS* was able to predict a wider variety of labels accurately. This is likely because the BMS covariates explain a significant amount of the variance in mean *ME* scores.

6.3. Task 3: Clustering limited patient data to find patterns

After reducing the original data (n = 24714) with Factor Analysis of Mixed Data to 300 dimensions, we can see the eigenvalues and explained inertia of the reduction in Table 5 in the Appendix. Since we will have 300 eigenvalues, I only display the largest 5 values. It is important to note the values are pretty small, thus no specific feature is contributing significantly to variance within the data.

For *UPDRS2*, we define n as the number of patient samples and k as the number of samples that contained *UPDRS2* scores. The histograms in Figure 10 and Figure 11 in the Appendix show the distribution of *UPDRS2* scores within each cluster. Cluster 0 (n = 2337, k = 52) was the only distribution that contained a good amount of 3 scores in *UPDRS2*. The main trend is seen within Cluster 1 (n = 7770, k = 32), Cluster 3 (n = 6611, k = 44), and Cluster 4 (n = 4400, k = 31) with mostly 0 and 1 scores. Lastly, Cluster 2 (n = 3596, k = 23) is similar to cluster 0, but with less 3 scores.

For *H&Y*, we again have n as the number of patient samples and k as the number of samples that contained *H&Y* scores. The histograms in Figure 12 and Figure 13 show the distribution of *H&Y* scores within each cluster. Cluster 0 (n = 6179, k = 45), Cluster 2 (n = 3993, k = 28), and Cluster 5 (n = 3385, k = 22) follow a similar distribution centered on score 2. Cluster 3 (n = 2776, k = 26), and Cluster 4 (n = 6497, k = 37) both contain a relative even distribution of scores. Lastly, Cluster 1 (n = 1884, k = 24) contains all 4 scores within the data.

7. Discussions

Task 1 showed that movement survey scores *ME* provide a good proxy for *UPDRS2*. It also showed that even using a very low number of training points it is possible to gain insight into a patients *H&Y* scores just using patient's survey answers. However, the number of samples does pose a serious limitation on the performance.

The results of Task 2 imply that it is possible to effectively estimate a patients *ME* (*UPDRS2* proxy) score based on limited data about a patient. One of the primary findings in the experimentation results was that augmenting a patient's Health and Demographic information with a quick survey of their motor functions (*BMS*) can significantly improve performance. This means that when predicting *UPDRS2* scores, while basic patient information seems to contribute meaningfully to prediction, a model meant to be applied in the field would likely need to include some basic survey data that assesses symptomatic characteristics for patients (as *BMS* does) in order to perform truly well. Future research directions might include improving the model to predict a wider variety of labels more consistently, perhaps by modifying the model's design or choosing additional simple survey covariates that contribute to label variance. Such an improved model could be used to fill out PD stage scores (ie. *UPDRS* scores) in datasets that lack them, and would enable certain datasets to be used for progression-related research.

The clustering results highlight that clustering with Factor Analysis of Mixed Data may not be the optimal method for assessing *UPDRS2* and *H&Y* scores from simple medical data as no clear patterns arise from the obtained clusters. Clustering methods should be further explored using Fox Insight to perhaps achieve other goals, such as PD subtyping or a similar method used in Doshi-Velez et al. (2014).

8. Acknowledgements & Member contributions

When deciding our approach for how to explore and gain insight from this dataset, Dr. Brett-Beaulieu Jones from Harvard Medical School helped guide us in the right direction. He helped educate us on the technicalities of PD, PD-related data, and PD-related research, helping us understand the importance of PD ratings scales and providing suggestions on the ML techniques that we could apply to achieve our exploration goals. Furthermore, Dr. Peter Szolovits from MIT CSAIL helped us formulate and optimize our approaches, and provided valuable counsel throughout the course of the project.

Member	**Worked on**
Gokul Kolady	Task 2 (Methods, Experiment Setup and Pre-Processing, Results), Discussion, Acknowledgements
Nicholas Ramirez	Task 3 (Methods, Experiment Setup and Pre-processing, Results), Related Work, Discussion
Marek Travnik	Task 1 (Methods, Experimental Setup, Results), Introduction, Related Work, Data Exploration, Discussion, Editing

Appendix

The code created for this project can be found at https://github.mit.edu/travnik/PD-progression.

Survey question	**Coef.**
Over the past week, have you had problems with your speech?	0.165
Over the past week, have you usually had problems with balance and walking?	0.126
Over the past week, have you usually had trouble turning over in bed?	0.113
Have you experienced falling in the last month?	0.103
Over the past week, have people usually had trouble reading your handwriting?	0.823

Table 4: The survey questions with highest positive importance in predicting Hoehn and Yahr scores using Lasso regression.

Largest Eigenvalues	Largest Explained Inertia
4.14496127e-05	0.05856058
1.67912012e-05	0.02372284
1.39442966e-05	0.01970069
1.31495933e-05	0.01857793
7.87245844e-06	0.01112232

Table 5: Largest eigenvals. and largest explained inertia.

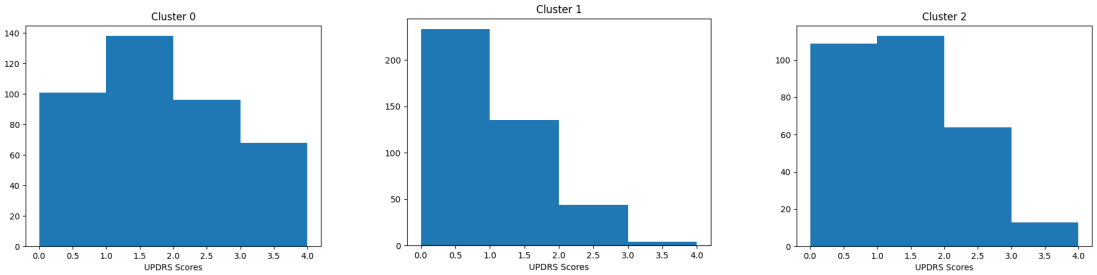

Figure 10: UPDRS2 Distribution in Clusters 0-2

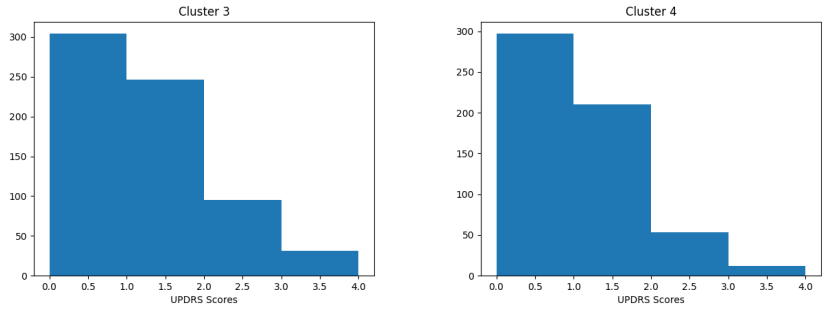

Figure 11: UPDRS2 Distribution in Clusters 3 and 4

Figure 12: H&Y Distribution in Clusters 0-2

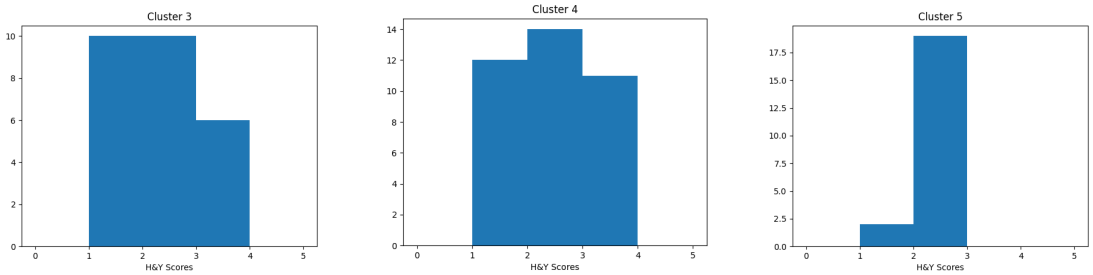

Figure 13: H&Y Distribution in Clusters 3-5

References

Finale Doshi-Velez, Yaorong Ge, and Isaac Kohane. Comorbidity clusters in autism spectrum disorders: An electronic health record time-series analysis. *Pediatrics*, 133(1):e54–e63, 2014. ISSN 0031-4005. doi: 10.1542/peds.2013-0819. URL https://pediatrics.aappublications.org/content/133/1/e54.

Alex Frid, Edmond J. Safra, Hananel Hazan, Lorrey L. Lokey, Dan Hilu, Larry Manevitz, Lorraine O. Ramig, and Shimon Sapir. Computational diagnosis of parkinson's disease directly from natural speech using machine learning techniques. In *2014 IEEE International Conference on Software Science, Technology and Engineering*, pages 50–53, 2014. doi: 10.1109/SWSTE.2014.17.

Christopher G. Goetz, Barbara C. Tilley, Stephanie R. Shaftman, Glenn T. Stebbins, Stanley Fahn, Pablo Martinez-Martin, Werner Poewe, Cristina Sampaio, Matthew B. Stern, Richard Dodel, and et al. Movement disorder society-sponsored revision of the unified parkinson's disease rating scale (mds-updrs): Scale presentation and clinimetric testing results: Mds-updrs: Clinimetric assessment. *Movement Disorders*, 23(15):2129–2170, Nov 2008. ISSN 08853185. doi: 10.1002/mds.22340.

C. Salvatore, A. Cerasa, I. Castiglioni, F. Gallivanone, A. Augimeri, M. Lopez, G. Arabia, M. Morelli, M.C. Gilardi, and A. Quattrone. Machine learning on brain mri data for differential diagnosis of parkinson's disease and progressive supranuclear palsy. *Journal of Neuroscience Methods*, 222:230–237, 2014. ISSN 0165-0270. doi: https://doi.org/10.1016/j.jneumeth.2013.11.016.

Sachin Shetty and Y. S. Rao. Svm based machine learning approach to identify parkinson's disease using gait analysis. In *2016 International Conference on Inventive Computation Technologies (ICICT)*, volume 2, pages 1–5, 2016. doi: 10.1109/INVENTIVE.2016.7824836.

Luba Smolensky, Ninad Amondikar, Karen Crawford, Scott Neu, Catherine M. Kopil, Margaret Daeschler, Lindsey Riley, Ethan Brown, Arthur W. Toga, Caroline Tanner, and et al. Fox insight collects online, longitudinal patient-reported outcomes and genetic data on parkinson's disease. *Scientific Data*, 7(1), 2020. doi: 10.1038/s41597-020-0401-2.