Modeling progression of Parkinson's disease

by

Christina X. Ji

B.S., Massachusetts Institute of Technology (2019)

Submitted to the Department of Electrical Engineering and Computer Science
in partial fulfillment of the requirements for the degree of

Master of Engineering in Electrical Engineering and Computer Science

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

September 2019

Author .....
Department of Electrical Engineering and Computer Science
August 2, 2019

Certified by .....
David A. Sontag
Associate Professor
Thesis Supervisor

Accepted by .....
Katrina LaCurts
Chairman, Department Committee on Graduate Theses

Modeling progression of Parkinson's disease

by

Christina X. Ji

Submitted to the Department of Electrical Engineering and Computer Science
on August 2, 2019, in partial fulfillment of the
requirements for the degree of
Master of Engineering in Electrical Engineering and Computer Science

Abstract

Chronic diseases progress slowly over years and impose a significant burden on patients. To help alleviate this burden, we propose tackling three clinical questions: predicting progression events, summarizing patient state, and identifying prognosis-driven subtypes. These questions are challenging because progression is highly heterogenous across patients. In this thesis, we address these challenges for Parkinson's disease (PD), the second-most common neurodegenerative disorder, using various machine learning approaches. First, we process data from the Parkinson's Progression Markers Initiative to convert it into a format that is easier to use for downstream machine learning analyses. Utilizing this data, we design novel data-driven outcomes that capture impairment in motor, cognitive, autonomic, psychiatric, and sleep symptoms and allow for heterogeneity in the patient population. Then, we build survival analysis models to predict these outcomes from baseline. Using our motor and hybrid outcomes can reduce the sample sizes and enrollment time for early PD clinical trials. We can provide further reductions by identifying more severe patients for enrollment via survival analysis and binary classification methods. For summarizing patient state, we seek better representations of disease burden by learning trajectories of disease progression. Lastly, we consider ways to use these patient representations and outcomes for discovering subtypes that capture differing rates of progression. We hope this thesis starts to answer the three clinical questions for PD and sparks more machine learning research in this area.

Thesis Supervisor: David A. Sontag
Title: Associate Professor

Acknowledgments

First and foremost, I would like to thank my advisor Prof. David Sontag. David constantly pushed me to focus on the most interesting questions, taught me how to see the big picture that makes our work meaningful, and ignited me with his contagious passion for research. I would also like to thank Rahul Krishnan for our numerous meetings and his mentorship. I was incredibly lucky to have had Suchan Vivatsethachai and Liyang (Sophie) Sun as my class project partners. They worked so hard on our class project that became sections 7.2 and 8.1 of this thesis. Our clinical collaborators Prof. Charles Venuto and Dr. Monica Javidnia (both University of Rochester) shared their invaluable clinical expertise with us, particularly in guiding our exploration of the PPMI data, sharing relevant papers, and making sure our methods and results were clinically sensible. We also had the privilege of working with Prof. Ray Dorsey and Prof. Karl Kieburtz (also University of Rochester) in defining the overall direction of our work.

I am fortunate to have had so many amazing research mentors leading up to this thesis project. In reverse chronological order: I would like to thank Dr. (and soon-to-be Prof.!) Fredrik Johansson for welcoming me into the lab and guiding me through the iterative process of reading papers and designing experiments. I learned a lot from Dr. Bin Liu (IBM Yorktown Heights) and Dr. Kenney Ng (IBM Cambridge) as they walked me through my first machine learning derivation. The serendipity of Dr. Alistair Johnson's UROP opening brought me into the machine learning for healthcare field, and I am grateful for his guidance in my first project in this area. I would also like to thank Prof. Roger Mark, Dr. Minnan Xu (Philips Cambridge), Dr. Junzi Dong (Philips Cambridge), Dr. Lauren Zasadil, Prof. Angelika Amon, Dr. Zhizhuo Zhang, Prof. Manolis Kellis, Dr. Jose Seoane Fernandez (Stanford), Dr. Ruping Sun (Stanford), Prof. Christina Curtis (Stanford), Dr. Phil Gedalanga (UCLA), and Prof. Shaily Mahendra (UCLA) for giving me the opportunity and guidance to get my toes wet in research.

I am also lucky to have had teachers who helped me develop my technical skills

and grow as a person. I would like to thank my academic advisor and instructor Prof. Pete Szolovits for sharing his perspective from many decades in the field. I learned so much from Prof. David Gifford, (newly minted!) Dr. Haoyang Zeng, Prof. Leslie Kaelbling, and Prof. Tomas Loreno Perez, who taught the two courses that introduced me to machine learning. I appreciate how my instructor Prof. Anna Mikusheva had her door open to me as I transitioned career interests. I am also grateful to my instructor Dr. (and soon-to-be Prof.!) Luca Spolaor for giving me advice and confidence for taking on graduate school. I am indebted to my middle and high school teachers who made it possible for me to come to MIT, especially Mr. Randy Lomas, Mr. Cy Ogle, and the late Mr. Eric Thiel (all Pleasanton, CA).

I would like to thank all the members of the Clinical Machine Learning group who make each day at work so pleasant: Monica Agrawal, Arjun Khandelwal, Irene Chen, Mike Oberst, Rebecca Peyser, Zeshan Hussain, Dr. Sanjat Kanjilal, Helen Zhou, and Hunter Lang, as well as those mentioned above. I look forward to continuing to learn from you! MIT would not have been complete without my college family Shang-yun (Maggie) Wu, Yun Boyer, and Menghua (Rachel) Wu. They made this journey so joyful and memorable. I am also blessed that my high school friend Yitian Zou (Pleasanton, CA) is always just a phone call away. In my 4 years at MIT, there have been an uncountable number of people with whom I have shared a happy (or sad) memory, so please forgive me for any omissions: simply because your name is not in these acknowledgments does not mean our time together is not in my heart.

Lastly, none of this would have been possible without the constant support and love of my parents. Thank you so much for being there for me through all the tribulations and triumphs of my MIT journey, all that came before, and all that the future will bring! With eternal gratitude and love, I dedicate this thesis to you.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Machine learning has become increasingly popular in studying chronic illnesses, with applications ranging from early prevention and diagnosis to predicting treatment effect and future prognosis. There are many challenges unique to healthcare that still need to be addressed [87]. For example, underlying disease conditions may not be directly observable or quantifiable in data, so identifying these latent variables is one challenge. Another is that these conditions are continuously evolving, but how they evolve varies from person to person. Our view of patients is incomplete, as records may be missing features and data is only collected at medical visits, which occur at discrete timestamps interspersed at irregular intervals. The observations may also be affected by hidden confounders, which may lead to incorrect conclusions if left unaccounted for. Pervasive throughout healthcare data is interindividual and interoccasion variability. The first refers to how having different clinicians, patients, and caregivers collect the data may lead to different labels for the same observation. The second refers to how the same person may label the same observation differently at different timepoints. In addition to these technical challenges, integrating domain knowledge is essential to every step of the modeling process: defining the task, selecting the covariates, parametrizing the model, interpreting the results for any new insights, and ensuring safety for deployment.

In spite of all these challenges, many advancements have been made in recent decades. The collection of large datasets through clinical trials and observational

studies has paved the way for more complete views of patients. Machine learning research in latent variable modeling has started tackling the issues of variability and hidden confounding. Burgeoning collaborations between machine learning researchers and clinicians has made the development of models that might potentially impact clinical practice possible.

Cancer and Alzheimer's disease have been at the forefront of this research, mostly driven by how widespread they are. We elect to focus on the less-often studied but still prominent Parkinson's disease (PD). This choice was driven in part by the recent collection of a large body of data. In this chapter, we introduce PD, some motivating clinical questions, and why we believe machine learning is a potent tool for tackling these questions. Then, we outline the remaining chapters of this thesis.

1.1 Introduction to Parkinson's disease and current clinical practice

Parkinson's disease is the second most common neurodegenerative disorder [61]. It is characterized by a significant loss of dopamine-producing neurons, especially in the substantia nigra region, which is responsible for motor and reward. The hallmark symptom of PD is poor motor control, evidenced in tremors, postural instability, and difficulty walking. The clinical definition of Parkinson's is the presence of bradykinesia (slowness of movement), as well as rest tremor and/or rigidity. Responsiveness to the standard motor medication levodopa is is a supportive criteria, while alternate parkinsonism diagnoses or severe motor impairment within five years are exclusionary criteria. Although these diagnostic criteria and hallmark characteristics are primarily motor symptoms, the effect on patients is much more extensive. Cognitive decline, depression, anxiety, difficulty sleeping, autonomic dysfunction, and poor sense of smell, to name a few, are all linked with Parkinson's. A reference timeline for when these symptoms start to appear is shown in Figure 1-1.

Fortunately, with the dopamine replacement treatment levodopa and other symp-

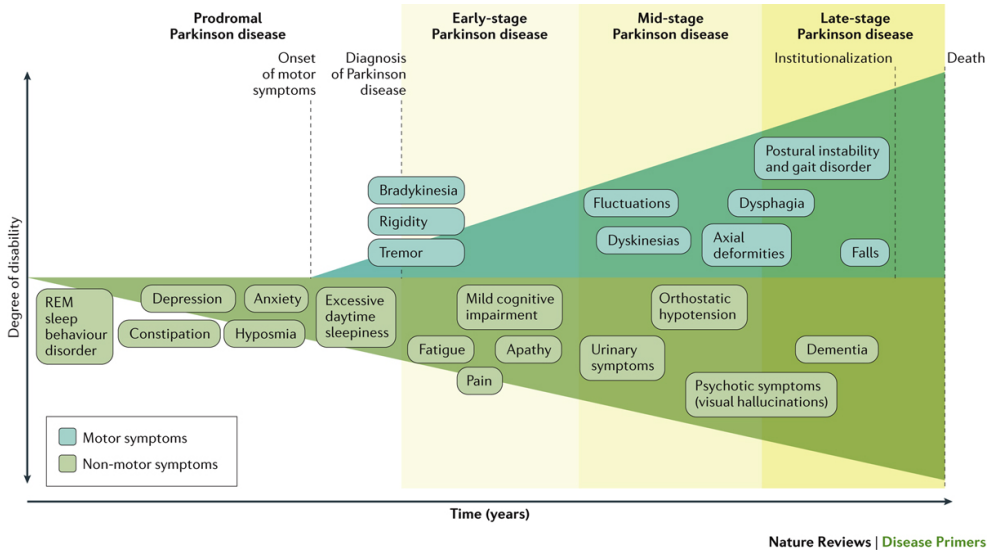

Figure 1-1: **Clinical symptoms associated with PD progression** from Poewe et al [61]. Hyposmia is reduced sense of smell. Bradykinesia is slowness of movement. Dyskinesias are involuntary movements. Dysphagia is trouble swallowing.

tomatic treatments, most symptoms can be managed, so patients can live for many decades without too much impact on quality of life. However, none of the treatments modify disease progression, so the disorder will eventually cause severe disability. Research on finding therapies that can delay progression is a high priority. To this end, understanding the multiple factors that cause Parkinson's disease and drive heterogeneity in progression and identifying biomarkers that can predict progression are critical.

1.2 Open clinical questions

Being diagnosed with a chronic illness can be daunting, as patients must consider how their future will be affected: Will their disease change quickly and make them have to give up their jobs and become reliant on a caregiver? Or can they manage the disease with treatments and continue their daily lives with only minor accommodations? For a PD diagnosis, the disease may not impose a significant burden until later, but patients would still like to know when they need to start making accommodations. Because of how heterogeneous the disease is, some patients may need aid walking,

while others will lose fine hand movements, and still others need to take sleeping aids or psychiatric medications or anticipate cognitive impairment. To understand and predict this heterogeneity, our goal is to explore three clinical questions: predicting progression events, summarizing and predicting patient state, and understanding subtypes of progression patterns. Below, we describe each question in more detail:

Predicting progression events: Confinement to a wheelchair, falling, or dementia are obvious markers of significant progression. However, these severe events are rarely observed and occur much later in the course of the disease. Instead, we want to identify intermediate events that already start to interfere with daily life. To formalize this, how do we characterize significant progression of Parkinson's disease in a way that allows for heterogeneity in progression? Can we predict when these progression events will happen for patients? Can we use delays in these events to measure treatment efficacy in clinical trials?

Summarizing patient state: For any illness, when a patient presents a set of symptoms, doctors need to get a bigger picture of what underlies these symptoms. This may seem like more of a question for diagnosis, such as when a patient enters the emergency room, but even after diagnosis, understanding changes in the underlying factors can shed light on future changes and symptoms. That is why we need a way to understand patient state. Even if we cannot observe the etiology, can we obtain some representation of overall disease severity that still accounts for heterogeneity?

Subtyping progression patterns: A multitude of PD subtyping systems have been developed, yet none have been integrated into clinical practice. Marras et al [47] identify two criteria clustering methods should satisfy to overcome this barrier: 1) A useful subtyping system should reflect at least one of the following: underlying biological causes, prognosis, or treatment responsiveness. 2) Patients should be easily classified into a single subtype so that doctors can actually apply the subtyping system. For the first criterion, the data we have is best equipped to study prognosis. Therefore, the question becomes: how can we identify subtypes that are informative of future progression? To address the second criterion, are there baseline features that can help us categorize patients into these subtypes?

1.3 Statement of contributions

In chapter 2, we introduce the dataset that we will be using: the Parkinson's Progression Markers Initiative (PPMI). We hope this delineation of the dataset, along with the processing scripts we provide to convert it into a format that is easier to use for downstream machine learning analyses, will make it easier for future researchers to use the dataset for tackling these clinical questions.

In chapter 3, we discuss some of the clinical work that has been done towards addressing these questions. We also review how machine learning has contributed to disease progression modeling and some specific papers that we build on.

As the main contribution of this thesis, we address the first clinical question of predicting progression events in chapters 4 through 6. As seen in Figure 1-2, we first define novel data-driven outcomes that are better suited for early-stage PD in chapter 4. In chapter 5, we predict these outcomes using survival analysis models. In chapter 6, we demonstrate how these outcomes and predictive models can be used to significantly reduce sample sizes required for clinical trials.

We take two approaches to start answering the second question of summarizing and predicting patient state in chapter 7. Both learn trajectories of patient states across time. The first directly learns the observed features, while the second uses unsupervised latent variable modeling.

In chapter 8, we consider three preliminary explorations of the last question of subtyping. The first builds on the latent variables from chapter 7, and the second considers another way to summarize the entire trajectory into a few components. For both of these, we make sure our subtyping methods are clinically useful by characterizing prognosis for each subtype and demonstrating how to classify patients into the subtypes from baseline. For the last exploration, we explore whether it is feasible to use the outcomes we define in chapter 4 for subtyping and leave the subtyping analysis for future work.

Finally, we conclude in chapter 9 with some thoughts on how our work has informed us of other interesting directions future researchers can take to better under-

stand these and other clinical questions related to Parkinson's disease.

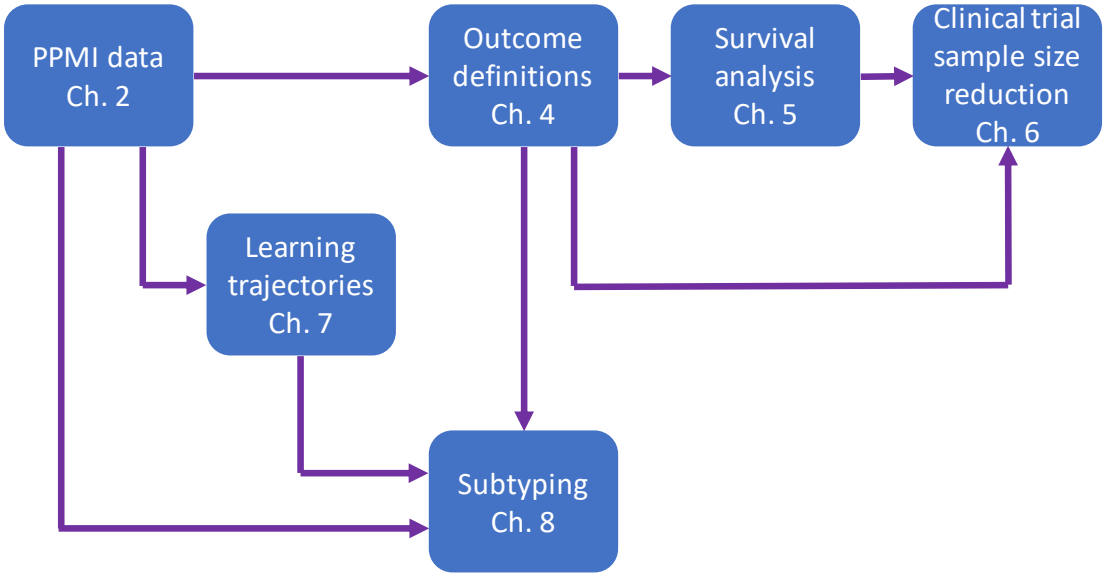

Figure 1-2: **Outline of this thesis.**

Chapter 2

Parkinson's Progression Markers Initiative Dataset

The Parkinson's Progression Markers Initiative (PPMI, `http://www.ppmi-info.org/`) is a landmark observational clinical study [46]. The study enrolled Parkinson's disease (PD) patients, healthy controls, and other relevant cohorts from sites across the US, Europe, Israel, and Australia from 2010 to 2018. The goal of PPMI is to help enable biomarker discover by providing a longitudinal record of clinical assessments, biomarkers, and imaging data.

In the remainder of this chapter, we give an overview of the patient cohorts in PPMI, the data collected, and how we processed it. We also summarize the availability for each cohort and give population-level statistics. For further information on the cohorts and assessments, please refer to the PPMI study documents [45].

We will provide our processing scripts in a repository on the ClinicalML GitHub page. Our scripts were run on data downloaded on January 24, 2019. The downloaded data contains a single file for each assessment. We combine these files into a few csvs for each cohort and perform some additional data cleaning. The output from our scripts is in an easier-to-use format for machine learning and other data science tasks. We hope that the scripts we provide, in conjunction with this chapter introducing the dataset, will help future researchers using the PPMI dataset and contribute to the search for more clinical insights.

2.1 Patient cohorts

The PPMI study sought to understand progression of PD by not only characterizing Parkinson's patients but also how they differ from other people [45]. For this purpose, they also tracked healthy control subjects and performed the same longitudinal assessments on them. In addition, they identified some common PD mutations and enrolled subjects who carried these mutations or had relatives with these mutations. These genetic carriers both with and without PD comprised a few of the other cohorts. Lastly, there has been interest in catching PD before diagnosis potentially to slow early neurodegeneration. To this end, the PPMI study also enrolled a cohort of high-risk patients before diagnosis. In this section, we summarize the eligibility criteria for each cohort of patients and introduce potential downstream use cases for machine learning studies.

2.1.1 *de novo* PD cohort

423 patients were enrolled in the *de novo* Parkinson's cohort. We refer to this group of patients as the PD cohort. The eligibility criteria for enrollment in this cohort were

- At least 2 of the following: resting tremor, bradykinesia, or rigidity; OR resting tremor or bradykinesia is asymmetric
- Diagnosis of Parkinson's within past 2 years and at age at least 30
- Hoehn and Yahr scale at stage 1 or 2
- DaTscan imaging showing dopamine deficit
- Not using PD medication within the past 60 days or expected to require PD medication in the next 6 months
- Not received any drugs that may interfere with DaTscan imaging
- Not receiving any treatment or having any condition that may interfere with lumbar puncture

Bradykinesia is slowness of movement. Hoehn and Yahr measures motor progression and will be detailed in Section 2.3.1. Lumbar puncture is the process used to collect cerebrospinal fluid, which will be explained in Section 2.4.3. Most work is focused on this group of patients, as they were given the most comprehensive assessments. Because they were all recently diagnosed, they are all in early stages of PD at enrollment.

2.1.2 Healthy controls

196 patients were enrolled in the healthy controls cohort, hereby abbreviated as the HC cohort. These patients were given most of the assessments prescribed to the PD cohort. As clinical studies rarely track healthy controls with such detail, the data collected on this HC cohort gives us the opportunity to contrast the PD cohort against the HC cohort and better understand how Parkinson's differs from normal aging. The eligibility criteria for enrollment in the HC cohort were

- Does not have neurological disorder or first degree relative with PD
- At least 30 years old
- MoCA score at least 26
- Last 2 criteria for PD cohort above

MoCA is a cognitive assessment we will introduce in Section 2.3.2, where a score below 26 indicates mild cognitive impairment.

2.1.3 Prodromal cohort

65 patients were enrolled in the prodromal cohort. Prodromal patients show early symptoms of PD. Usually these are nonmotor symptoms and as such, they do not have the motor symptoms that qualify them for PD diagnosis. The eligibility criteria for enrollment in the prodromal cohort were

- At least one of the following:

  - Hyposmia (loss of smell): defined as UPSIT score at or below the 10th percentile by age and gender
  - Diagnosis of REM sleep behavior disorder
- Age 60 or older
- GDS score at most 4 or at most 9 at investigator's discretion
- STAI score at most 53 or with investigator's discretion
- No diagnosis of dementia or PD
- Last 2 criteria for PD cohort

The smell assessment is described in Section 2.3.4, sleep in Section 2.3.5, and GDS and STAI in Section 2.3.3. In particular, we note from the second bullet that patients in this cohort are older. The clinicians we worked with also noted that because of the focus on specific nonmotor symptoms, patients enrolled in this cohort might be a subtype of Parkinson's.

At each visit, the patients in this cohort are given a prodromal diagnostic questionnaire. This records the clinician's assessment of the patient's diagnosis: no neurological disorder, motor prodromal, nonmotor prodromal, and Parkinson's disease. Nonmotor prodromal seems like the most common diagnosis. Of the 65 patients, 30 were nonmotor prodromal for the entirety of the study. 11 patients go from nonmotor prodromal to motor prodromal, and 8 patients oscillate between nonmotor and motor prodromal. 11 patients have no neurological diagnosis at some point. **Phenoconversion** is defined as the transition from prodromal or no diagnosis to being diagnosed with PD. This happened for 15 patients, with 3 going back and forth between phenoconversion and remaining in the prodromal phase. As this diagnosis is somewhat probabilistic, patients often oscillate between visits.

Because most of the dopaminergic neuron loss occurs before Parkinson's diagnosis, to alter the disease course, neuroprotective action must be taken before diagnosis. That makes understanding the prodromal cohort a high priority. Although we do not

take on these questions in this thesis, we discuss some future directions that came up in discussions with our clinical collaborators in chapter 9.

2.1.4 Genetic PD cohort

265 patients were enrolled in the genetic PD cohort, hereby referred to as GENPD. These patients may have more advanced Parkinson's disease. Less data is collected from them. The eligibility criteria for the GENPD cohort were

- First and last criteria for PD cohort
- Diagnosed with Parkinson's at most 7 years ago
- Hoehn and Yahr scale less than 4
- At least 18 years old
- Confirmation of causative LRRK2, GBA, or SNCA mutation
- No diagnosis of dementia
- Not received any drugs that may interfere with PET imaging

Background on the genetic factors is provided in Section 2.4.1.

2.1.5 Genetic unaffected cohort

399 patients were enrolled in the genetic unaffected cohort, abbreviated as GENUN going forward. These patients do not have Parkinson's disease but either carry or have first degree relatives who carry genes associated with PD. Less data is collected for them. The eligibility criteria for the GENUN cohort were

- At least one of the following:
  - At least 45 years old and either carrying or having a first degree relative who carries a LRRK2 or GBA mutation

  - At least 30 years old and either carrying or having a first degree relative who carries SNCA mutation

- No diagnosis of PD
- Last 2 criteria of GENPD
- GDS and STAI criteria for prodromal cohort

The younger age requirement for the SNCA gene is likely because SNCA is associated early-onset PD around age 50 [52].

2.1.6 Registry cohorts: PD and unaffected

203 and 249 patients were enrolled in the registry and unaffected PD cohorts, REGPD and REGUN for short, respectively. Fewer criteria are required for the registry cohorts, and even less data is collected for them. The eligibility criteria were

- At least 18 years old
- Carries or has a first degree relative who carries a SNCA mutation
- No diagnosis of dementia

PPMI collected the genetic and registry cohorts to study the effect of specific biomarkers. Some patients in these cohorts have had Parkinson's for a long time. If future research can leverage the use of a few timepoints much farther along in the Parkinson's trajectory, these cohorts can be an invaluable resource.

2.1.7 SWEDD cohort

SWEDD stands for scans without evidence for dopaminergic deficit. These patients exhibit the same symptoms as Parkinson's disease patients but do not have dopamine deficits in their DaTscan imaging. 64 patients were enrolled in this cohort. As such, they have the same eligibility criteria as the PD cohort except their DaTscan images shows no dopamine deficit. In the past, these patients were diagnosed with PD

because the outward symptoms are identical. However, it is becoming increasingly obvious that this may be a different disease altogether, so characterizing how these patients differ is another interesting clinical question.

2.2 Demographics

The following demographics were collected for all cohorts:

- Birthdate
- Date of consent to study, i.e. date of enrollment
- Study site number
- Gender: male or female
- Race: White, Asian, Black or African American, American Indian or Alaskan Native, Native Hawaiian or Pacific Islander, other (multiple allowed)
- Ethnicity: identify as being Hispanic or Latino (Spanish origin)
- Number of years of education
- Dominant hand: right, left, or mixed

These baseline features related to Parkinson's were also collected for all cohorts:

- Date of diagnosis if diagnosed
- Dominant side of disease if diagnosed: right, left, or mixed
- Family history of PD: biological mother, biological father, full sibling, half sibling, maternal grandparent, paternal grandparent, maternal aunt/uncle, paternal aunt/uncle, children

Cohort-level statistics for some of the features above are shown in Table 2.1.

A general medical history exam indicates if a patient has ever received a diagnosis in any of the following systems: dermatological, opthalmological, ear/nose/throat,

pulmonary, cardiovascular, gastrointestinal, hepatobiliary, renal, gynecologic/urologic, musculoskeletal, metabolic/endocrine, hemato/lymphatic, neurologic (other than PD), psychiatric, allergy/immunologic, and other.

Cohort	Age	Male	White	Fam hist	Time since diag
PD	61.6 (9.7)	65.5%	94.8%	12.8%	0.577 (0.530)
HC	63.1 (11.0)	64.1%	93.8%	1.0%	
Prodromal	68.8 (5.7)	78.5%	90.8%	9.2%	
GENPD	62.1 (10.3)	48.7%	97.4%	43.4%	3.234 (2.213)
GENUN	61.7 (7.7)	39.6%	98.5%	72.2%	
REGPD	70.3 (9.9)	53.5%	97.5%	41.9%	9.320 (6.361)
REGUN	49.4 (14.7)	39.1%	98.8%	68.3%	
SWEDD	60.7 (10.1)	62.5%	95.3%	25.0%	0.525 (0.517)

Table 2.1: **Cohort demographics at baseline.** Percentages are shown for male, white, and presence of family history of PD. Mean followed by standard deviation in parentheses are shown for the other features. Time since diagnosis is in years.

2.3 Clinical assessments

Parkinson's patients have five main categories of symptoms: motor, cognitive, autonomic, psychiatric, and sleep. Questionnaires designed to measure aspects of these symptoms are the primary way clinicians assess patients. Questions on these assessments are typically binary or have a numerical scale indicating degree of severity. Assessments are taken at the following times since enrollment (in years): 0, 0.125, 0.375, 0.625, 0.875, 1.125, 1.625, 2.125, 2.625, 3.125, 3.625, 4.125, 4.625, 5.125, 6.125, and 7.125. The average number of visits and enrollment time for each cohort are shown in Table 2.2. In the remainder of this section, we will introduce the assessments one-by-one. For the more commonly used assessments, we show some cohort-level trends and individual trajectories.

2.3.1 Motor assessments

MDS-UPDRS (Motor Disorders Society-Unified Parkinson's Disease Rating Scale) was designed to measure disease severity and is the most common outcome used in

Cohort	# visits	Time enrolled
PD	13.1 (3.2)	5.36 (1.71)
HC	5.8 (1.7)	5.47 (1.91)
Prodromal	10.9 (2.4)	3.75 (0.97)
GENPD	5.7 (2.8)	2.27 (1.43)
GENUN	4.4 (2.5)	2.21 (1.38)
REGPD	1.5 (0.7)	1.03 (1.20)
REGUN	1.3 (0.5)	0.72 (1.08)
SWEDD	7.7 (2.0)	2.24 (1.10)

Table 2.2: **Amount of longitudinal data.** # visits is the number of visits with an MDS-UPDRS exam. Enrollment time is in years. Standard deviation in parentheses.

clinical trials [58]. MDS-UPDRS is constructed of four components: part 1 primarily measures non-motor symptoms, part 2 is a self-assessment of motor symptoms in daily life, part 3 is a clinician assessment of specific motor tasks, and part 4 measures complications of dopaminergic therapy. Each MDS-UPDRS question is on a discrete scale from 0 to 4: 0=normal, 1=slight, 2=mild, 3=moderate, and 4=severe.

Part 3 of the MDS-UPDRS can be administered off medication or on medication. A patient is considered off medication if medication is withheld the night prior to testing. On and off designations only matter if a patient is taking dopamine replacement or a dopamine agonist. Patients on MAO-B inhibitors do not have on or off exams. We define an exam as on medication if it is specified as such or the medication time was specified to be within 6 hours of exam time. Otherwise, the exam is considered off medication. Averages across time for the PD, prodromal, and HC cohorts are shown in Figure 2-1.

In our analyses, we split up MDS-UPDRS parts II and III into subtotals based on clinician guidance and an examination of correlation between features, as seen in the heatmap in Figure 2-2. The subtotal definitions are shown in Table 2.3. Plots of the subtotals for 5 randomly selected PD patients are shown in Figure 2-3.

Patients are designated as **tremor dominant (TD)** versus **postural instability gait dominant (PIGD)** based on the MDS-UPDRS score [78]. The tremor score is defined as the mean of the 11 tremor variables in parts II and III. The postural instability score is defined as the mean of the questions on trouble walking and freezing

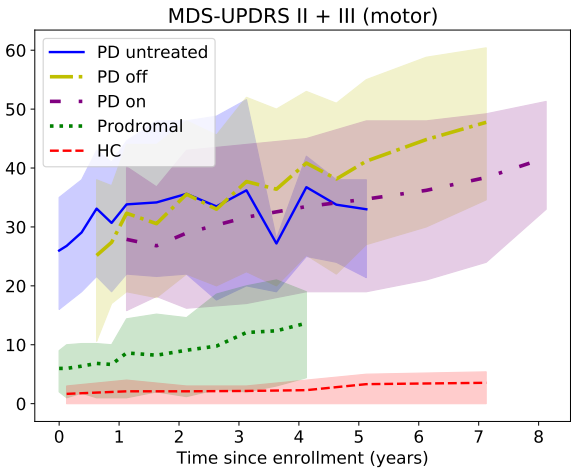

Figure 2-1: **MDS-UPDRS averages across time** for PD, prodromal, and HC cohorts by time since enrollment. Shaded regions are from 20th to 80th percentile.

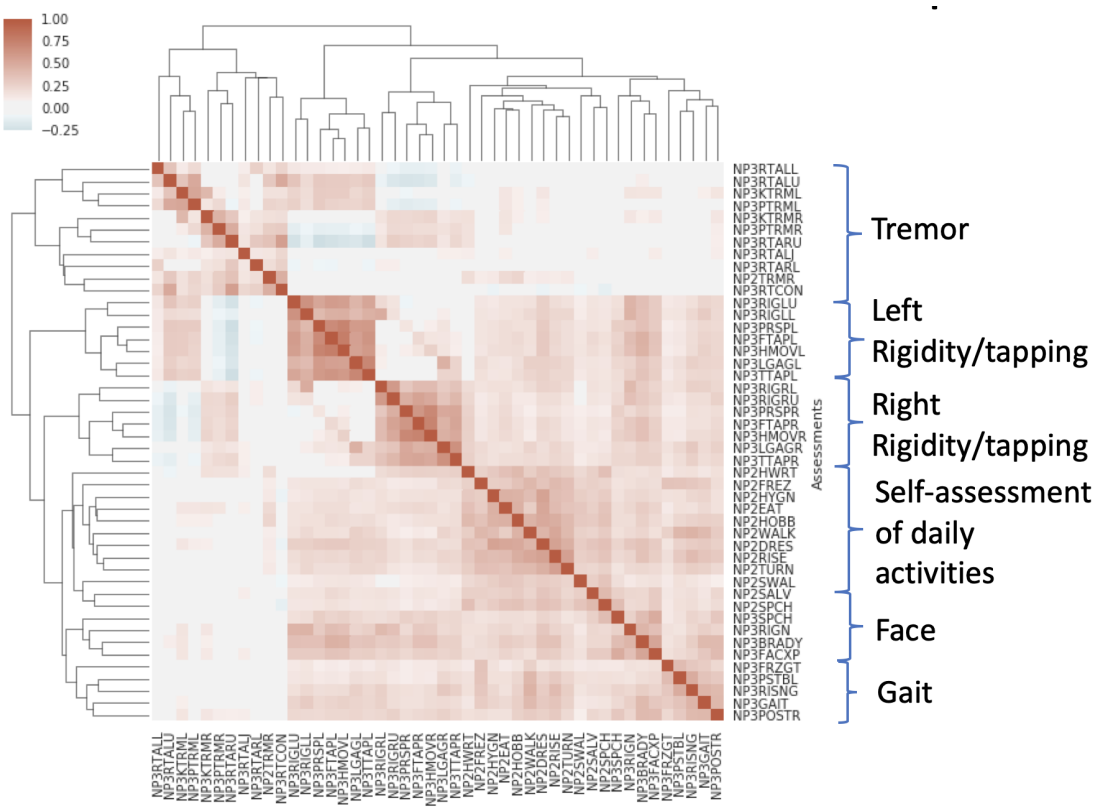

Figure 2-2: **Correlation heatmap of MDS-UPDRS questions** with subtotal annotations on the right.

Subtotal	MDS-UPDRS question numbers	Max
Daily activities	All questions in part II except 2.10	44
Face	3.1, 3.2, 3.3 neck, 3.14	16
Right rigidity	3.3 (2 questions), 3.4, 3.5, 3.6, 3.7, 3.8	28
Left rigidity	same as above but questions on left side	28
Tremor	2.10, 3.15 (2 questions), 3.16 (2 questions), 3.17 (5 questions), 3.18	44
Gait	3.9, 3.10, 3.11, 3.12, 3.13	20

Table 2.3: **Definitions of MDS-UPDRS subtotals**. Maximum is the total number of points that fall under that subtotal (4 points per question).

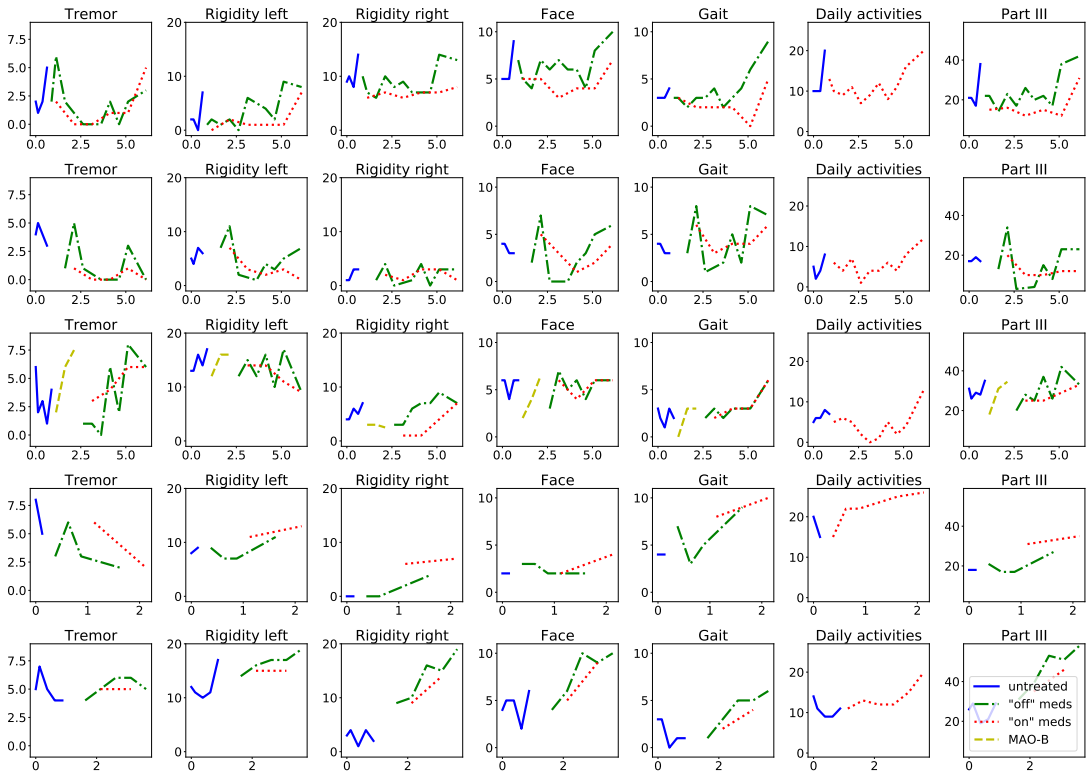

Figure 2-3: **MDS-UPDRS subtotals for 5 PD patients**

Stage	Description	Yr 0 freq	Yr 3 freq
1	Unilateral impairment with minimal disability	48.9%	10.2%
2	Bilateral impairment without affecting balance	51.1%	81.9%
3	Bilateral impairment with impaired postural reflexes and moderate disability	0.0%	5.7%
4	Severely disabled but still able to walk or stand unassisted	0.0%	1.3%
5	Confined to bed or wheelchair unless aided	0.0%	1.0%

Table 2.4: **Hoehn and Yahr** stage descriptions and percentages in PD cohort at year 0 and year 3.

of gait in part 2 and gait disturbances, freezing of gait, and postural instability in part 3. If the ratio of tremor to postural instability score is at least 1.15 or the postural instability score is 0 while the tremor score is nonzero, the subject is tremor dominant. If the ratio is at most 0.9, the subject is PIGD. Otherwise, the patient is indeterminate. In the PD cohort, 82.7% of patients are TD, and 5.7% of patients are PIGD at baseline. At year 3, 71.0% of patients are TD, and 32.0% are PIGD. Of these, 7.1% are TD from an MDS-UPDRS in 1 treatment setting and PIGD from an MDS-UPDRS exam in another treatment setting.

The **Hoehn and Yahr** (H&Y) scale is a general summary of motor features [28]. Table 2.4 describes the H&Y stages [20]. The majority of PPMI subjects are in the earlier stages.

Diagnostic features asks about risk factors for PD or stroke, atypical symptoms, tremor, rigidity, bradykinesia (slow movement), akinesia (loss of voluntary movement), postural or gait disturbances (including whether the patient is likely to fall), body hemiatrophy (asymmetric symptoms), and hyperkinesias (involuntary spasms). The survey also briefly covers psychiatric, cognitive, autonomic, and other features.

2.3.2 Cognitive assessments

MoCA (Montreal cognitive assessment) measures cognitive function, with questions on visuospatial ability (e.g. draw a clock, name animals), memory (remembering words or sequences), attention (e.g. subtraction), language (word- or sentence-level

fluency), and orientation (current time and location) [55]. The assessment is on a 30-point scale, where higher scores indicate better cognitive function. One additional point is given if a patient has fewer than 12 years of education. A score of at least 26 indicates the patient is cognitively normal.

Figure 2-4 shows some cohort-level averages for MoCA. Note that the prodromal cohort has a lower average MoCA score when plotted against time since enrollment since the eligibility criteria for the prodromal cohort requires they are older. The two cohorts have similar MoCA scores when plotted against age. MoCA for 5 example PD patients is shown in Figure 2-7.

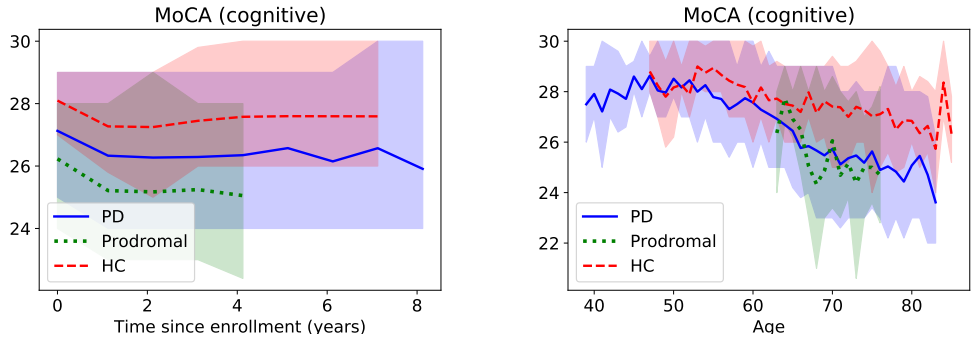

Figure 2-4: **MoCA averages** for PD, prodromal, and HC cohorts by time since enrollment on left and age on right. Shaded regions are from 20th to 80th percentile.

For the **HVLT** (Hopkins verbal learning test), the examinee is read 12 words and given 3 trials to recall all 12 words [5]. The sum of the 3 trials is the immediate recall score. Then, the examinee is read 24 words, consisting of the original 12 words, 6 related distractors, and 6 unrelated distractors. The patient must identify which 12 words were from the original set. Discrimination is defined as the number of true positive words recognized minus the number of false positive words recognized (related and unrelated). 20 minutes later, the patient is asked to again recall the 12 words. The retention score is the number of words recalled 20 minutes later divided by the maximum number of words recalled in the second or third trial originally.

For the **LNS** (letter number sequencing) test, the examinee is read a sequence of letters and numbers and asked to recall the numbers in ascending order and the letters in alphabetical order [89]. **BJLO** (Benton judgment of line orientation) asks

Cohort	HVLT retent	HVLT recog	HVLT recall	LNS
PD	0.855 (0.201)	9.634 (2.623)	24.443 (4.973)	10.585 (2.653)
HC	0.905 (0.180)	10.066 (2.789)	26.046 (4.486)	10.867 (2.564)
Prodromal	0.769 (0.290)	9.333 (2.204)	21.794 (5.280)	9.444 (2.844)
GENPD	0.814 (0.251)	9.946 (2.375)	23.950 (5.538)	9.704 (3.137)
GENUN	0.886 (0.167)	10.503 (1.892)	26.539 (5.039)	11.179 (2.886)
SWEDD	0.835 (0.230)	8.484 (3.767)	24.281 (4.343)	9.875 (2.637)
Cohort	**BJLO**	**Sem Fluency**	**SDM**	
PD	12.770 (2.128)	48.666 (11.621)	44.991 (9.175)	
HC	13.122 (1.978)	51.796 (11.173)	50.152 (10.336)	
Prodromal	11.968 (2.279)	45.000 (10.775)	44.308 (9.558)	
GENPD	11.564 (2.982)	48.523 (13.435)	42.840 (11.258)	
GENUN	11.179 (2.886)	55.990 (12.688)	49.526 (9.502)	
SWEDD	12.750 (2.359)	45.203 (12.327)	45.404 (10.976)	

Table 2.5: **Cognitive at baseline**. Mean followed by standard deviation in parentheses for each assessment across cohorts. Cognitive categorization assessment indicated normal function for all patients at screening. The two registry cohorts have no samples.

the examinee to match two angled lines to a set of lines that are arranged in a semicircle and separated 18 degrees from each other [3]. **SDM** (symbol-digit modalities) measures speed of cognitive processing by asking patients to translate a symbolic code into a sequence of numbers when given the legend [75]. The **semantic fluency** test asks the examinee to say as many words that belong in the following categories as possible in one minute: animals, vegetables, and fruits [27]. On the **cognitive categorization** assessment, the clinician is asked to assess whether the patient has normal cognition, mild cognitive impairment (MCI), or dementia and give a level of confidence about this assessment. We only take the measurement if the confidence level is at least 50%. **MDS-UPDRS question 1.1** also measures cognitive impairment. Cohort-level statistics for these exams at baseline are shown in Table 2.5.

2.3.3 Psychiatric assessments

There are 3 primary psychiatric illnesses that are assessed: depression, anxiety, and impulsive-compulsive behavior. **GDS** (Geriatric depression scale) measures the first on a 15-point scale [73]. A score of at least 5 on GDS indicates a patient has depres-

sion. **STAI** (state-trait anxiety inventory) consists of 40 points for state anxiety and 40 points for trait anxiety [77]. State anxiety refers to anxiety induced by a particular situation. Trait anxiety is associated with daily life. The two are highly correlated, as seen in Figure 2-5. The Pearson r coefficient is 0.789. STAI for 5 example patients is shown in Figure 2-7. **QUIP** (questionnaire for impulsive-compulsive behavior for PD) measures if a patient feels like they cannot control their behavior in the past 4 weeks related to gambling, sex, buying, eating, or other activities [90]. Some questions in **MDS-UPDRS part I** also measure psychiatric symptoms: 1.2 asks about hallucinations and psychosis, 1.3 about depression, 1.4 about anxiety, 1.5 about apathy, and 1.6 about compulsive behavior (also known as dopamine dysregulation syndrome since it could be a side effect of dopaminergic medications). Cohort-level statistics for the first 3 questionnaires are shown in Table 2.6. Cohort-level averages across time are shown in Figure 2-6.

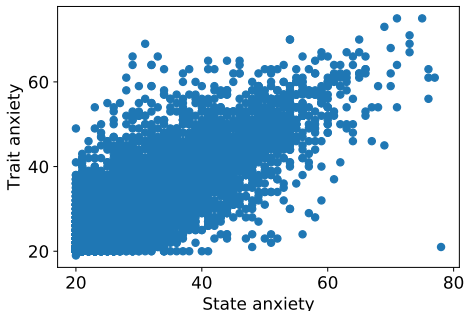

Figure 2-5: **State and trait anxiety correlation**

2.3.4 Autonomic assessments

SCOPA-AUT measures whether a patient has had trouble with swallowing, constipation, urine retention, lightheadedness, excessive sweating, tolerating light/cold/heat, and sexual activity [86]. All questions are on a 0 to 3 scale measuring how often the symptoms occurred in the past month. **UPSIT** (UPenn smell identification test) measures if a patient has reduced sense of smell, a common early symptom of PD [13]. This symptom is called **hyposmia**. **MDS-UPDRS part I** measures autonomic

Cohort	GDS yr 0	GDS yr 3	STAI yr 0	STAI yr 3
PD	14.0%	16.9%	65.4 (18.3)	64.9 (18.7)
HC	6.6%	4.2%	57.2 (14.0)	55.5 (13.5)
Prodromal	12.3%	27.8%	63.1 (17.1)	64.6 (19.0)
GENPD	29.1%	34.5%	71.4 (17.8)	60.4 (14.9)
GENUN	11.1%	9.1%	61.8 (17.8)	60.4 (14.9)
SWEDD	29.7%		69.8 (18.0)	
Cohort	**QUIP yr 0**	**QUIP yr 3**	**# ppl yr 0**	**# ppl yr 3**
PD	0.0%	23.2%	423	366
HC	0.0%	16.2%	196	167
Prodromal	0.0%	16.7%	65	53
GENPD	0.0%	35.2%	265	91
GENUN	0.0%	20.6%	399	68
SWEDD	0.0%		64	

Table 2.6: **Psychiatric at baseline**. Percentages with depression and impulsive disorders. Mean followed by standard deviation in parentheses for STAI. The two registry cohorts and SWEDD at year 3 are omitted since at most 2 people were observed for each.

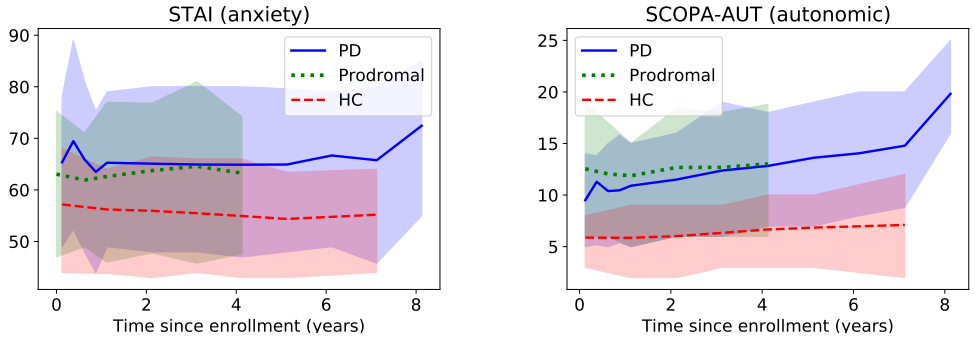

Figure 2-6: **STAI and SCOPA-AUT averages across time** for PD, prodromal, and HC cohorts by time since enrollment. Shaded regions are from 20th to 80th percentile.

Cohort	SCOPA-AUT yr 0	SCOPA-AUT yr 3	UPSIT yr 0
PD	9.499 (6.146)	12.380 (7.015)	22.352 (8.220)
HC	5.877 (3.733)	6.335 (4.031)	33.980 (4.845)
Prodromal	12.547 (7.521)	12.685 (6.727)	17.180 (6.505)
GENPD	12.977 (8.137)	15.910 (9.557)	21.637 (8.869)
GENUN	8.352 (6.100)	9.167 (6.473)	33.115 (5.176)
REGPD			20.650 (5.535)
REGUN			33.089 (5.535)
SWEDD	13.750 (8.773)		31.359 (6.178)

Table 2.7: **Autonomic at baseline**. Mean followed by standard deviation in parentheses for autonomic and smell assessments. Refer to the table for psychiatric features for roughly the number of people at each timepoint. Missing values have too few samples.

function in questions 1.9 through 1.13. In addition to the general areas covered by SCOPA-AUT, MDS-UPDRS also measures pain and fatigue. Cohort-level statistics for the first 2 assessments are shown in Table 2.7. Cohort-level averages across time for SCOPA-AUT are shown in Figure 2-6. SCOPA-AUT for 5 example patients is shown in Figure 2-7.

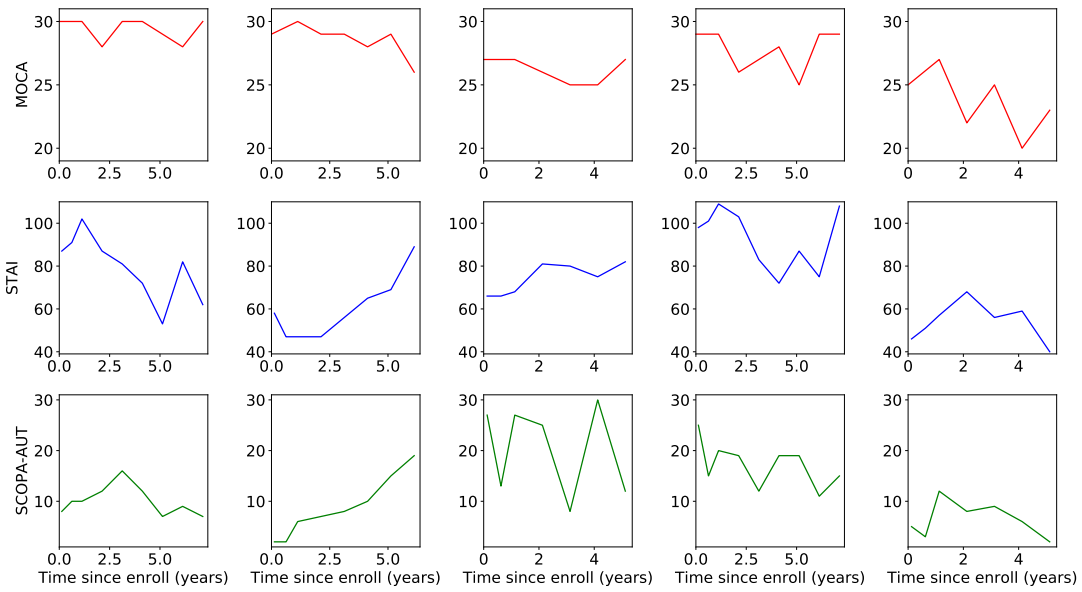

Figure 2-7: **MoCA, STAI, and SCOPA-AUT for 5 PD patients**

Cohort	Epworth yr 0	Epworth yr 3	REM yr 0	REM yr 3
PD	15.6%	27.5%	37.6%	43.7%
HC	12.3%	15.6%	19.9%	16.2%
Prodromal	21.9%	13.0%	70.3%	61.1%
GENPD	23.2%	29.2%	40.2%	30.3%
GENUN	10.0%	6.1%	20.8%	15.2%
SWEDD	32.8%		40.6%	

Table 2.8: **Percentages with sleeping disorders**. Refer to the table for psychiatric features for roughly the number of people at each timepoint.

2.3.5 Sleep assessments

ESS (Epworth sleepiness scale) measures how likely a patient is to doze off during certain activities (from watching TV to talking to someone) during the day on a scale of 0 to 3 [34]. A total of at least 10 indicates excessive daytime sleepiness. **REM sleep behavior disorder** measures nighttime sleep disturbances, such as vivid dreams that involve physically moving the body [79]. It also records whether a patient has had a neurological disorder in the past as those are risk factors for REM sleep behavior disorder. A total of at least 5 indicates a patient has REM sleep behavior disorder. **Questions 1.7 and 1.8 in MDS-UPDRS** measure nighttime sleep issues and daytime sleepiness, respectively.

2.3.6 Miscellaneous assessments

A **neurological exam** measures the 10 cranial nerves, muscle strength, coordination, sensation, muscle stretch reflexes, and plantar reflex (flexor, extensor, or indeterminate when foot is stimulated by a blunt instrument).

PASE (physical activity scale for the elderly) has two components: household and leisure [88]. This survey asks patients what activities they engaged in (e.g. walking, sports, housework, and work), how often, and for how many hours per day. **MSEADLG** (Modified Schwab and England activities of daily living) asks patients to rate themselves on a score ranging from totally dependent at 0 to completely independent at 100 [70]. Most subjects in PPMI are in the upper end of the range, so this assessment is not very informative.

On the **physical exam**, a clinician indicates if the patient has any abnormalities with skin, head/neck/lymphatic, eyes, ears/nose/throat, lungs, cardiovascular system, abdomen, musculoskeletal system, neurological system, psychiatric, or other. The following **vital signs** are measured: weight, height, temperature, supine systolic and diastolic blood pressures, standing systolic and diastolic blood pressures, and supine and standing heart rates.

2.4 Biomarkers

2.4.1 SNPs

Single nucleotide polymorphisms (SNPs) are the most common type of genetic variation among people. They vary from having no effect to causing diseases. In total, 208 SNP alleles are reported in the PPMI dataset [25, 26]. We restrict the set of SNPs we use to those that are measured for at least 90% of patients and have a minor allele frequency of at least 5%. This set includes 146 SNPs. Whole genome sequencing data is also available if a researcher would like to extract more genes. These are the 3 most commonly studied genes: The **GBA** gene produces a protein that is involved with the cellular disposal process in lysosomes [61]. Parkinson's patients with a mutation in the GBA gene often manifest at a younger age and show more significant cognitive impairment. **SNCA** is also known to be associated with early-onset Parkinson's disease. Over 30 mutations in SNCA have been identified [61]. The SNCA gene produces alpha-synuclein, which is measured as a CSF biomarker and discussed in Section 2.4.3. **LRRK2**, another gene related to Parkinson's disease, is potentially associated with protein trafficking and lysosomes [61]. In addition, **COMT** encodes the enzyme catechol-O-methyltransferase, which breaks down dopamine and other neurotransmitters [74]. Nalls et al [54] calculate a **genetic risk score** for Parkinson's disease using 90 alleles [31].

2.4.2 RNA expression

Gene expression data from three studies are included in the PPMI dataset. Potashkin et al [65, 64] ran real-time qPCR (quantitative polymerase chain reaction) on RNA collected from blood to measure expression levels. They used the following set of primers: **COPZ1, ZNF160, PTBP1, C5ORF4, WLS, SOD2, APP, HNF4A, and EFTUD2**. GAPDH was included as a housekeeping gene. The data is given in the unit Ct, which is the number of cycles required to reach a detectable threshold.

To normalize this data, we use the $\Delta\Delta$Ct method [18]. The first step is to take the average of the housekeeping genes and then subtract the average from each of the genes of interest for each patient. This gives the ΔCt. Next, take the average of the ΔCt for each gene across the healthy controls cohort and subtract this average from the ΔCt to obtain the desired $\Delta\Delta$Ct.

As detailed in Dotan et al [12], real-time qPCR with the following set of primers: **HSPA8, SKP1, PSMC4, ALDH1A1, UBE2K, and LAMB2**. GAPDH and PGK1 were used as the housekeeping gene. We took the average of the 2 replications and performed the same $\Delta\Delta$Ct normalization method described above.

Scherzer et al [69] run a slightly different qPCR procedure to count the number of transcripts of each splicing variant of a gene. Alternative splicing is when parts of the RNA are cut out and the remainder is put together. Different splicings lead to different functions. The following genes are measured: **ZNF746, DHPR, SNCA-007, SNCA-3UTR-2, SNCA-3UTR-1, SNCA-E4E6, SNCA-E3E4, FBX07-001, FBX07-005, FBX07-007, FBX07-008, FBX07-010, DJ-1, and GLT25D1**. The housekeeping genes they use are SRCAP, RLP13, UBC, MON1B, and GUSB. FBX07 is associated with early-onset PD.

To normalize this data, we follow Anders et al [1] and first calculate the geometric mean for each patient of the counts of all housekeeping genes. Then we divide the gene counts by this geometric mean and take the median across all patients for each gene. This median is called the size factor of a gene. The transcript count divided by this size factor is the desired quantity.

2.4.3 CSF biomarkers

Cerebrospinal fluid (CSF) is collected from a patient via a lumbar spinal puncture. The following biomarkers are measured using this sample:

- **Abeta-42**, a 42-amino acid protein, which aggregates to form senile plaques.
- **total Tau** and **pTau** (Tau phosphorylated at threonine-181): Tau proteins become defective and can no longer stabilize microtubules properly in neurodegenerative diseases, instead forming neofibrillary tangles.
- **Alpha-synuclein**: function unclear but potentially related to presynaptic terminals.

Following recommendations by Coffey et al [7] in the PPMI derived variables specifications, we also look at the ratios between Abeta-42, pTau, and total Tau. The distributions of these derived features at baseline are shown in Figure 2-8

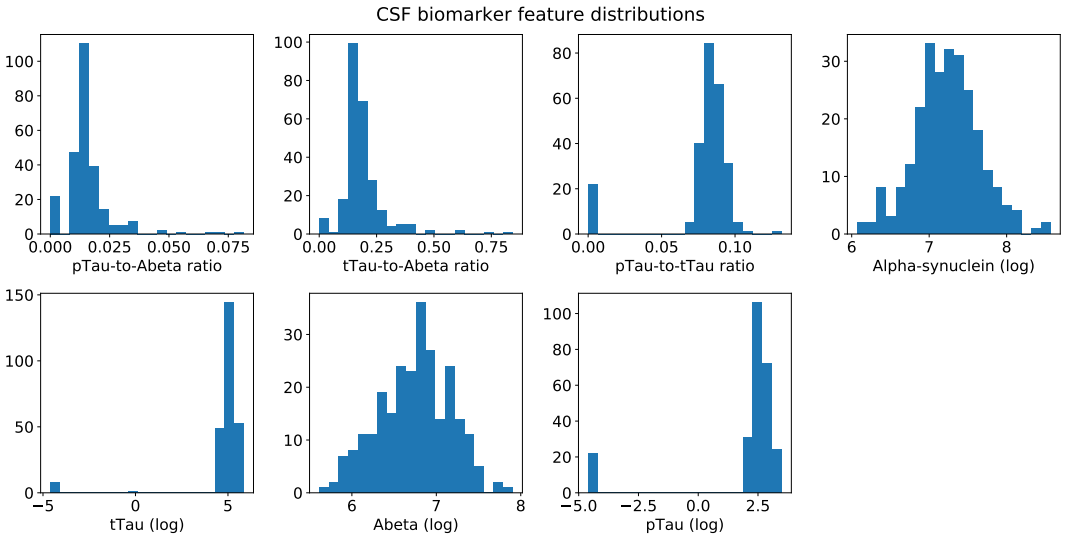

Figure 2-8: **CSF biomarker distributions at baseline**.

2.4.4 Blood hematology

A standard lab test is run on blood samples. The biomarkers measured include **alanine aminotransferase (ALT)**, **aspartate aminotransferase (AST)**, and

alkaline phosphatase. These enzymes are high in the blood if the liver is damaged. **Prothrombin time** measures the extrinsic pathway in coagulation. **APTT** (activated partial thromboplastin time) measures the intrinsic pathway. For the **immune system**, lymphocytes, monocytes, WBC, platelets, eosinophils, neutrophils, and basophils are counted. Whether the red blood cells show any deformed shapes is also recorded. Chloride, uric acid, potassium, sodium, glucose, and bicarbonate are measured using the serum. Total protein, creatinine, hematocrit, and hemoglobin are also recorded. We take the SI unit measurement if it is available and convert from the US unit if it is not. The blood hematology biomarkers are flagged if they are high or low.

2.4.5 Other biomarkers

Some biomarkers that are occasionally taken include the **neurotransmitters** dopamine, serotonin, noradrenaline, and adrenaline, as well as some of their metabolites. Several types of **sphingomyelins** (the membrane sheath surrounding axons in neurons) are also measured.

2.5 Imaging

2.5.1 DaTscan imaging

DaTscan is a dopamine transporter (DaT) single photon emission computerized tomography (SPECT) imaging technique that uses the radioactive imaging agent ioflupane [23]. This method is primarily used to measure dopamine levels in the brain. Image analysis is primarily focused on the caudate and putamen regions, which comprise the dorsal striatum of the basal ganglia and regulate motor control, the reward system, and some forms of learning. PPMI processed the images to compute the dopamine levels on the left and right side of the caudate and putamen. These levels are recorded as the striatal binding ratio (SBR).

Following the PPMI derived variables specifications [7], instead of left and right,

contralateral and ipsilateral are typically used as features instead. If the dominant side of the body affected by Parkinson's is the right side, then contralateral is left and ipsilateral is right. If the dominant side is mixed or the patient is a healthy control, then the average of the two sides are taken for both contralateral and ipsilateral. Other features derived following PPMI specifications include mean caudate, mean putamen, count density ratios (of caudate to putamen), and asymmetry indices. The distributions at baseline are shown in Figure 2-9.

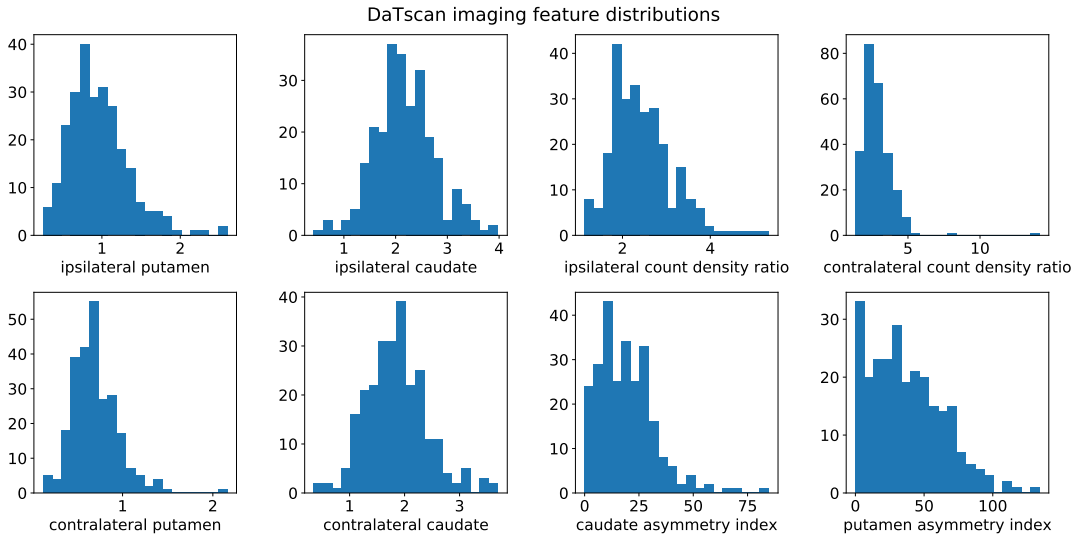

Figure 2-9: **DaTscan imaging feature distributions at baseline**.

2.5.2 Other imaging modalities

Other imaging modalities that are available include **MRI** (magnetic resonance imaging), **DTI** (diffusion tensor imaging), **PET** (positron emission tomography), and **functional MRI**. For DTI, 6 regions of interest and 2 reference regions were identified in the PPMI dataset. The average of rows were taken if they shared the same run date. The availability of raw images in these modalities are shown in Table 2.9. Some example images are shown in Figure 2-10. With convolutional neural networks and more recent advances in imaging analysis, these brain images could provide more insights on PD.

Cohort/Image type	Year 0	Year 1	Year 2	Year 4
PD DaTscan	414	366	342	193
PD DTI	157	147	133	105
PD MRI	393	151	133	103
Prodromal DaTscan	0	1	0	0
Prodromal DTI	18	15	15	0
Prodromal MRI	213	92	33	8
HC DaTscan	177	22	1	0
HC DTI	69	64	13	14
HC MRI	182	64	12	14

Table 2.9: **Number of patients with various imaging modalities** available in years 0, 1, 2, and 4. Images were rarely available during other years.

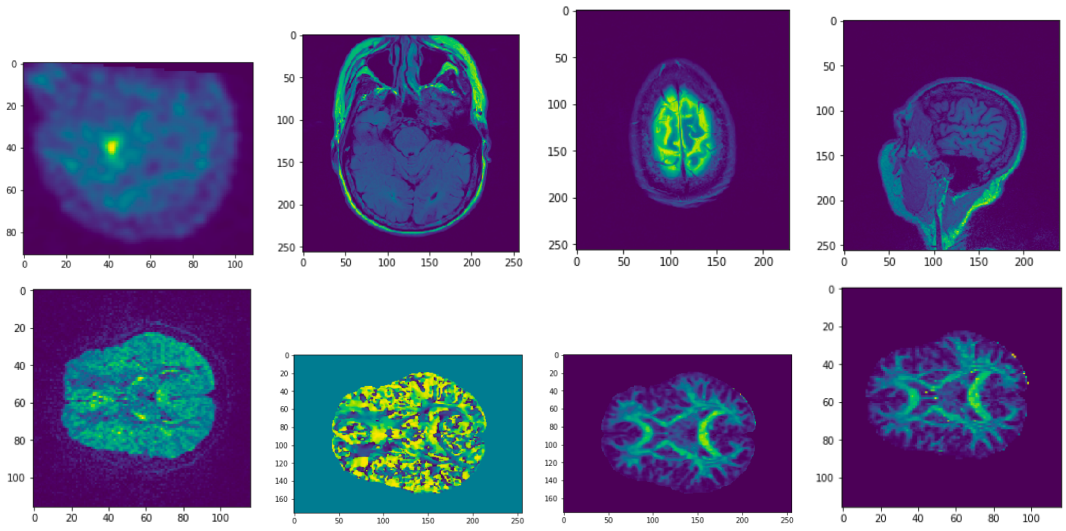

Figure 2-10: Examples of **imaging modalities**. Left to right: Top row: DaTscan, MRI axial fluid-attenuated inversion recovery, MRI axial turbo spin echo, MRI saggital magnetization-prepared rapid gradient echo. Bottom row: DTI 4-d motion trajectory, DTI eigenvectors of MRI, DTI fractional anistrophy of MRI, DTI fractional anistrophy of EPI.

2.6 Treatments

PPMI logs when a patient started and stopped taking a medication, the number of times a medication is administered per day, and the dosage per administration. All medications are recorded regardless of whether they are directly related to PD.

PD medications typically act on the dopamine pathway and are primarily prescribed to treat motor symptoms. The three main ways they act on the pathway are shown in Figure 2-11: 1) supplying more dopamine, 2) activating dopamine receptors, and 3) reducing dopamine breakdown.

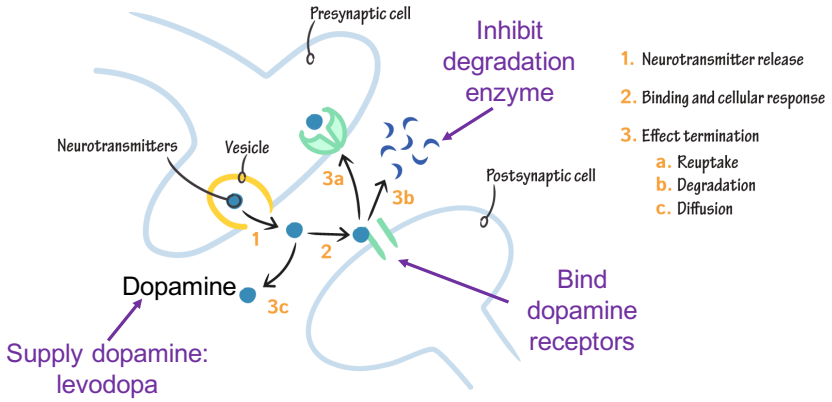

Figure 2-11: **Three mechanisms for dopaminergic treatment**.

Levodopa is the primary Parkinson's medication [21]. It works according the first mechanism above as a precursor to the neurotransmitter dopamine. Taking levodopa alone results in nausea, so it is always coupled with carbidopa. Carbidopa helps ensure that most of the levodopa is not converted to dopamine until after it crosses the blood-brain barrier since dopamine cannot cross the barrier. This significantly reduces the dosage of levodopa required. Levodopa is also available in slow-releasing forms, usually with a COMT inhibitor such as entacapone. A side effect of levodopa is spontaneous movement, known as dyskinesia. Patients also experience off-periods when the medication suddenly wears off. The effectiveness and side effects of levodopa use are recorded using MDS-UPDRS part IV.

Dopamine agonists are another class of motor medications that work by acti-

vating dopamine receptors. Examples include pramipexole, ropinirole, and rotigotine. **MAO-B** (monoamine oxidase B) is an enzyme that breaks down dopamine. Medications that inhibit this enzyme include selegiline and rasagiline. Other Parkinson's medications include amantadine, apomorphine, propanolol, and anti-cholinergics.

Conversions to levodopa-equivalent daily dosages are provided by PPMI [29]. These medications are often taken in combination, as seen in Figure 2-12, which was gathered using the MDS-UPDRS assessment data in PPMI. The average time to treatment initiation is 1.4 years since enrollment, with a standard deviation of 0.8 years.

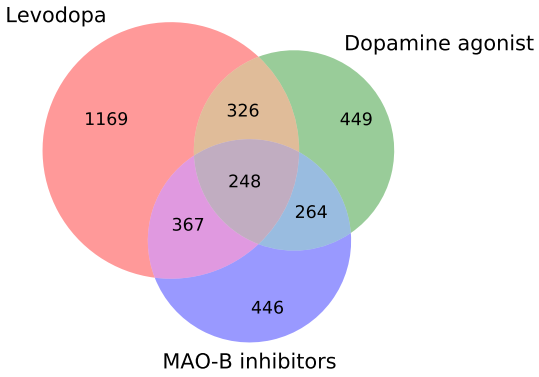

Figure 2-12: **Treatment usage** at each patient visit. Not depicted: 2,255 visits were untreated.

Using the drug name to class mappings provided by our clinical collaborator Monica Javidnia, we identify the drug classes shown in Table A.1. Classes 1-11 are more closely associated with the assessments detailed previously, so classes 12-18 may be omitted from some of our analyses. We only record binary indicators for whether a patient is on treatment rather than the dosage. Dosage is hard to calculate because there are many types of units and the name of the drug would need to be parsed for the ratio between two drugs in a combined pill. Based on the start and end dates indicated for each drug, we align with the 'INFODT' field for visits to determine whether a patient is taking a drug in that class on a given visit. If a patient receives a treatment within 3 months of the initial visit, we record that treatment as taken

at the initial visit. For later visits, any treatment that was started after the previous visit are included in the following visit. If dates are missing, we assume the treatment was taken once on the entry date.

2.7 Output from our data processing

The assessment files were merged using 'PATNO' and 'EVENT_ID.' 'INFODT' was used if 'EVENT_ID' was no available in a file. Time since enrollment was assigned based on 'EVENT_ID' and the PPMI schedule of activities. The screening visit occurs on the consent date, and the difference between the consent date and diagnosis date was used to calculated the time since diagnosis. For symptomatic therapy (ST) visits, which are unplanned, they are matched to their corresponding visits using the 'INFODT' field [80]. We only use data from patients who were enrolled at some point in PPMI.

As output from our data-processing scripts, we produce 7 csv files for each cohort:

- Features that are only measured at the screening visit: SNPs, medical history, and physical exam
- Features that are only measured at the baseline visit: demographics and UPSIT
- Assessment totals that are measured longitudinally
- Assessment questions that are measured longitudinally
- Other longitudinal features
- Treatments that were started more than 3 months prior to the first visit
- Treatments at each visit

Our files can easily be read in using pandas [51] and converted into numpy arrays for downstream analysis [83].

Data used in the preparation of this article were obtained from the Parkinson's Progression Markers Initiative (PPMI) database (`www.ppmi-info.org/data`). For

up-to-date information on the study, visit www.ppmi-info.org. PPMI-–a public-private partnership-–is funded by the Michael J. Fox Foundation for Parkinson's Research and funding partners, including abbvie, Allergan, Avid, Biogen, BioLegend, Bristol-Myers Squibb, Celgene, Denali, GE Healthcare, Genentech, GlaxoSmithKline, Lilly, Lundbeck, Merck, Meso Scale Discovery, Pfizer, Piramal, Prevail, Roche, Sanofi Genzyme, Servier, Takeda, Teva, UCB, Verily, Voyager Therapeutics, and Golub Capital.

Chapter 3

Related work

In this chapter, we will introduce relevant papers on disease progression modeling. In particular, we describe related work in the context of the three clinical questions outlined in the introduction, including work specific to Parkinson's and papers on machine learning methodology that we adopt.

3.1 Disease progression modeling

Disease progression modeling studies the dynamics by which a disease evolves in a patient and how treatment affects these dynamics. This field originated from pharmacodynamics. Post et al [62] describe models on a biological level, where treatment effect is modeled as a disturbance to a cellular or tissue system. Cook et al [8] provide a review of where the field stands as of 2016. On the opposite end of the spectrum from the biological models described by Post et al are empirical models. Purely data-driven, they decompose the process into baseline disease status, rate of disease progression over time, and treatment effect. Then, each part is a function parametrized by covariates, such as demographics or baseline features. Linear, exponential, inverse polynomials, more complex compositions are chosen based on trial observations. Venuto et al [85] review some empirical models that have been developed for Parkinson's disease. Most of this work has been centered on the MDS-UPDRS score. For the disease progression part, a logistic curve is a popular choice since

growth becomes bounded. For the treatment effect part, these models often reflect how it takes some time for a drug to start taking effect. Once full effect is reached, the difference between treated and untreated trajectories is mostly a constant offset since treatments do not alter progression.

3.2 Machine learning for disease progression modeling

Machine learning researchers primarily focus on empirical models because of the data that is available to them. The main disctinction between machine learning and pharmacodynamic models is the use of latent variables. Instead of directly modeling how an assessment changes over time, the assessments are treated as observed outputs from a latent variable. The models then capture how this latent variable evolves across time and how the assessment can be emitted from this latent variable. The models often take on more complex forms to account for heterogeneity and explicitly model noise in the emission function. Some models are learned using Bayesian inference, Markov chain Monte Carlo sampling, expectation-maximization, or more complex learning algorithms.

For example, Xu et al [92] use a 3-part decomposition similar to Post et al [62] for their model and add a noise component. Schulam et al [72] further develop this model by coupling features to learn correlations across trajectories of different features. Soleimani et al [76] consider heterogeneity in the disease progression component by splitting it into a population-level function and a patient-specific function. Having both parts allows for some statistical power to be gained from the whole population. Schulam et al [71] add another level of resolution to this model at the subpopulation level. They also structure the noise component so that it specificallly captures transient trends rather than interpatient variability.

Wang et al [87] developed an unsupervised disease progression model, where they assign a meaning to each latent variable using anchors. Anchors are observed features

that have the latent variable as their sole parent. In their model, the latent axes are the onset of co-morbidities, and the anchors are specific ICD-9 codes related to those co-morbidities. The remaining features are children of all the latent variables and can be used to infer any of them. To allow for varying time intervals, they define generator matrices that represent transition rates rather than transition probabilities. The transition probabilities can then be calculated by multiplying the rate by the time interval and taking a matrix exponential.

These models may also be useful for the Parkinson's disease setting, but because we have a limited number of patients in PPMI and do not focus on symptomatic treatment effect, we do not adopt these models here. We do, however, consider other latent variable models, to which we give a basic introduction in section 3.5.

3.3 Predicting progression

Predicting rates of change or future scores has been commonly studied in Parkinson's disease. For example, Latourelle et al [37] study motor progression using the MDS-UPDRS exam. Specifically, they fit a linear regression to the on and off exam scores to get 1 slope for each of the treatment settings for each patient. They then predict the slopes from 18 baseline demographics and assessments, 8 DaTscan imaging features, 7 CSF biomarkers, and 17,456 SNPs. In particular, they identify SNPs that might be associated with rapid motor progression. For example, they found that patients who carried minor alleles of rs17710829 and rs929897 have substantially faster rates.

Mollenhauer et al [53] is another study that predicts annual rates of change using baseline features. In particular, they are interested in comparing *de novo* Parkinson's patients and healthy controls, in terms of motor and cognitive progression, measured using MDS-UPDRS part III and the Mini Mental State Examination (MMSE) [82], respectively. As such, they implement a 2-group latent curve model, which includes a latent intercept factor to account for different baseline severity and a latent slope factor to account for different annual rates of change. Then, they identify which factors were predictive of the latent slope for PD but not for HC. Our work could also

benefit from such comparisons to try to disentangle the effects of PD versus aging.

3.4 Clinical outcome measures

Most work on studying outcomes has been done in the context of clinical trials. Before a trial, the investigators define what outcomes they want to measure in patients. If the trial demonstrates that the proposed treatment reduces the number of patients who reach an adverse outcome, then the treatment might be considered effective. Oftentimes, these outcomes are biomarker measurements. Sometimes, they are based on when treatment is required, when motor complications start to arise, or changes in MDS-UPDRS score. See Parkinson Study Group [22] and Biglan et al [4] for examples. Marras et al [48] review how outcomes to clinical trials should be associated with patient quality of life. They state that standard outcomes in use do not meet this criterion. In the absence of new validated patient assessments, we consider how to use the assessments we currently have to develop better outcomes. Ellis et al [15] investigate this question using an alternative quality of life assessment known as Parkinson's Disease Questionnaire-39 (PDQ-39) [33]. They found that demographic factors, particularly disease duration and age, accounted for 19.7% of the variance in the PDQ-39 score. Within motor assessments, bradykinesia contributed a significant portion of variance, while tremor and rigidity did not.

De Pablo-Fernández et al [10] collect nonstandard retrospective data from PD patients who have died. The data includes some demographics, levodopa response level, and severe outcomes including falls, use of a wheelchair, dementia, and home care. They classify patients into the three subtypes diffuse malignant, mainly motor-slow progression, and intermediate from Ferenshtehnejad et al [17] (see section 3.6 for description). Then, they show the subtypes clearly stratify the survival curves for time from diagnosis to these outcomes.

Hughes et al [30] track 83 PD patients and 50 healthy controls over 10 years to study the onset of dementia. However, only 17 patients fulfilled the criteria for dementia over the 10 years. They found that male sex, age at entry into the study, and

severity of motor symptoms also at entry were predictive of onset of dementia, while duration of PD and age at onset of PD were not. With a more intermediate cognitive outcome, we posit that more informative factors could potentially be discovered.

McGhee et al [50] consider an outcome for morbidity (poor patient health) or mortality. They develop a new outcome called "dead or dependent," where dependent refers to a score below 80 on the the modified Schwab & England assessment. This outcome satisfies the criterion of Marras et al [48]. McGhee et al demonstrate its potential efficacy as a clinical trial outcome. We find that early-stage PD patients rarely encounter this outcome though. Macleod et al [43] develop some survival models for predicting this outcome. As baseline covariates, they consider age, male sex, postural and gait instability, MMSE score, and the Charlson co-morbidity index [66]. For more information on survival analysis models, refer to section 5.1. For our work with this outcome in particular, see chapter 6.

3.5 Latent variable modeling

Factor analysis is a method for empirically looking for underlying structure among observed features [81]. The entities in the underlying structure are called factors. Let n be the number of samples, d the number of features, and f the number of factors. Then let the $n \times d$ matrix X be the data, the $1 \times d$ matrix μ be the means, the $n \times f$ matrix F be the factors, the $f \times d$ matrix L be the loading matrix that maps from factors to features, and the $n \times d$ matrix ϵ be noise. Factor analysis solves for L and F in the equation $X - \mu = LF + \epsilon$ with the following constraints: F and ϵ are independent, $E[F] = 0$, and $Cov(F) = I$ to ensure the factors are uncorrelated. The learned factors are then evaluated to see if they represent something interpretable. Common use cases of factor models include reducing dimensionality and denoising. This serves as a baseline for more complex latent variable models.

An autoencoder is an artificial neural network widely used to perform dimensionality reduction. It consists of two parts: an encoder that maps the inputs to low-dimensional latent factors and a decoder that reconstructs the inputs from the

latent factors. A commonly used variant is the variational autoencoder (VAE) [36, 11]. A VAE is tailored towards generating new samples rather than reconstructing ones it has already seen. Samples are encoded as probability distributions. In addition to minimizing the reconstruction loss, the Kullback-Leibler divergence (distance between the encoding probability distributions) is also minimized so that the encoding space is somewhat compact, and interpolation can be performed between samples that the model has seen. Compared to factor analysis, VAE is more adept at handling non-linearities with neural networks and disentangling to make the latent factors more distinct.

Pierson et al [60] build on the VAE framework to infer multi-dimensional rates of aging from cross-sectional data. They model the latent variable as evolving linearly over time. Then, they apply a nonlinear emission function to map the latent variables to the observed features. They constrain this nonlinear function to be monotonic since health tends to deteriorate as people age but also include a noise component. They then prove that with this constraint, the rates are identifiable and interpretable. They also use a small collection of follow-up visits to demonstrate their model can extrapolate to future timepoints. We apply this model and extensions of it to PPMI in section 7.2.

3.6 Subtyping

Many PD subtyping systems have been proposed. In particular, motor and nonmotor assessments at baseline or some later timepoint are typically used as the features in k-means clustering. This method learns k cluster centers, where each sample is assigned to the nearest center, so that the sum of squared distances from each sample to its cluster center is minimized. This is typically learned by alternatively assigning samples to their nearest clusters and updating the center to be the mean of the new assigned samples until convergence.

van Rooden et al [84] review 7 studies that were performed before 2010. They found that having a cluster for rapid progression and older onset age and another

cluster for slow progression and younger onset age was common. However, these designations differed in terms of proportion of PD population and degree of cognitive impairment across different study definitions. Marras et al [47] perform another review of 9 subtyping studies until 2012. They find similar levels of disagreement across studies. In addition, they highlight some reasons these subtyping systems have not been implemented in practice: 1) They are not related to underlying biology, prognosis, or treatment response. 2) They are not intuitive to use, especially as some patients may fall into multiple clusters due to ambivalent boundaries.

Fereshtehnejad et al [17] review subtypes of PD as of 2017. They mention how the earliest PD subtyping systems were based solely on age of onset and motor symptoms, but with data-driven methods, clustering now includes nonmotor symptoms, prognosis, and biomarkers. For example, one system that has been defined is diffuse malignant (severe motor and nonmotor symptoms), mainly motor-slow progression, and intermediate (everything else). One challenge they highlighted was understanding when and why patients might change between subtypes.

An example subtyping system by Lawton et al [38] is shown in Figure 3-1. In the two inner rings, they distinguish the clusters by rate of motor progression and levodopa response. In the two outer rings, they profile the baseline features. Interestingly, cluster 3 seems to have the most severe baseline symptoms but only intermediate progression. One problem with this system is that patients might not fall neatly into one of the clusters. For example, if a patient progresses slowly and has good levodopa response, would they belong in cluster 2 or 4? Likewise, if at baseline, a patient has more gait and posture problems, symmetric motor problems, and average or better-than-average nonmotor symptoms, the patient might not fall into any of the clusters.

Faghri et al [16] take a slightly different approach to subtyping. They start with dimensionality reduction on the assessment questions from all timepoints, specifically non-negative matrix factorization. Due to the correlation structure, the matrix factors neatly into motor, cognitive, and sleep dimensions. They verify the three latent factors reflect progression by applying the learned loading matrix to the healthy control

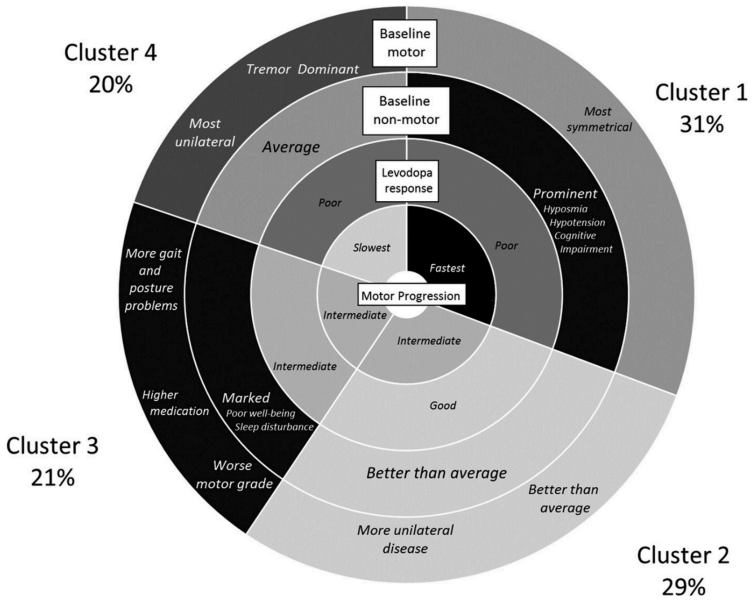

Figure 3-1: **Salient clinical features of the four clusters** from Lawton et al [38]. The extra 1% is due to rounding.

population and demonstrating that the latent factors progress away from healthy controls over time. With these three latent factors, they fit a Gaussian mixture model to identify three clusters of Parkinson's patients. We explore their methods in more detail in section 8.2.

There have been two studies applying deep learning to PD subtyping. The first uses a time-aware long short term memory (LSTM) network [2]. An LSTM is a type of recurrent neural network that consists of a memory cell and input, forget, and output gates. By time-aware, they mean the model accounts for how much time elapses between consecutive visits. To do this, they decompose the memory cell into a long-term and a short-term component. Then, they apply a time decay to the short-term component. To learn a representation of the entire patient trajectory, they stack two time-aware LSTMs: one as an encoder and another as a decoder. Then, they use this representation for k-means clustering and examine the feature profiles associated with each cluster.

The second paper also uses an LSTM [93], an extension of their earlier paper [6]. However, instead of stacking two LSTMs to learn a representation of the entire trajectory, they apply dynamic time warping to measure the distance between the

trajectory of hidden states of two patients. Lastly, they apply t-SNE to embed the trajectories into two dimensions and perform k-means clustering again (refer to section 8.3 for a description of the t-SNE method). One of the 3 subtypes they discovered represented particularly severe PD patients. We elect not to use LSTM or other deep learning models because of sample size and interpretability concerns.

Chapter 4

Outcome definitions

In healthcare, mortality is a common concern, whether someone has an acute or a chronic disease. For Parkinson's disease, in which mortality may not occur for decades after diagnosis, an occurrence or delay in mortality is not an efficient measure to demonstrate a delay in disease progression. Instead, morbidity–a state of poor health–is a more pressing issue. This can be represented by increases in disease severity over shorter time intervals. For example, predicting disease severity as assessed by the MDS-UPDRS scale at a future timepoint is a common task. Another is to estimate rate of change as the slope of a linear regression or the difference between two timepoints. This rate can also be predicted from baseline features. See Latourelle et al [37] for an example. However, as we can see from the visualizations of assessments for individual patients (Figures 2-3 and 2-7), the assessments are noisy and subject to interindividual and interoccasion variability. As such, models developed to predict future assessment values or rates of change may be affected by noise.

That is why we propose defining binary outcomes. Binary outcomes are less susceptible to noise because we can define an outcome as sustained severity. Since PD is characterized by motor and nonmotor symptoms, we define a separate outcome for each category of symptoms: motor, cognitive, psychiatric, autonomic, and sleep. The category outcomes can then be combined to form a hybrid outcome that measures how symptoms associated with PD affect daily life. In this chapter, we outline the design criteria and outcome definitions. In chapters 5 and 6, we illustrate a couple use cases

of these outcomes as prediction targets for survival analysis and alternative outcomes for clinical trials. We hope this work will spark discussion for other applications and ways to define clinically relevant outcomes.

4.1 Definition of outcomes related to symptom severity

As introduced in chapter 2, each category has assessments that are designed specifically to measure those symptoms. For motor and autonomic function, the definitive assessments are MDS-UPDRS and SCOPA-AUT, respectively. For cognitive performance, a wide array of assessments have been developed, and seven are commonly measured in PPMI. Psychiatric symptoms commonly present in three ways: depression, anxiety, and compulsive behavior, the last largely due to dopaminergic therapies to replace loss of dopamine cells. Each is measured by a different assessment. Likewise, there are two primary types of sleep disorders: excessive daytime sleepiness measured by the Epworth assessment and nighttime problems measured by the REM sleep behavior disorder questionnaire. We synthesize the diverse assessments collected in PPMI into a single outcome.

4.1.1 Design criteria

Our outcomes are defined hierarchically: first, we set the threshold for each assessment independently. An assessment must be above the threshold if higher scores indicate more severe symptoms. We try out possible thresholds starting from the 30th percentile and incrementing by every 5th percentile for continuous features. The search is stopped at the first threshold that satisfies the criteria. If lower scores indicate more severe symptoms, an assessment must be below the threshold. A similar search starting from the 70th percentile and decrementing is performed. Then, we set the number of assessments required for a category outcome, testing from one upwards. Lastly, we set the number of category outcomes required for the hybrid outcome.

This process is depicted in Figure 4-1. For all three steps, we use the same criteria as follows:

1. It takes at least 3 years for more than 50% of the Parkinson's cohort to cross the threshold.
2. At least 5% of the PD cohort cross the threshold.
3. At least 50% of the PD cohort must have a measurement after 3 years.

The workflow for using these three criteria to set the threshold for a single feature or category is shown in Figure 4-2. The purpose of the first criterion is so that the outcome would not occur too early in advance. We select 3 years and 50% by manually inspecting the survival curves. Because clinical trials are limited in time frame by cost, the outcome needs to occur early enough for a significant number of patients so that it can be detected. However, if the outcome is too early, it would not be as clinically meaningful since it does not signal sufficient progression. The second and third criteria are to avoid having significant censoring. Although we only used the PD cohort with the selection criteria, we also looked at the survival curves for the other cohorts for validation, as we would generally expect that healthy controls should have much fewer and later outcomes compared to PD patients.

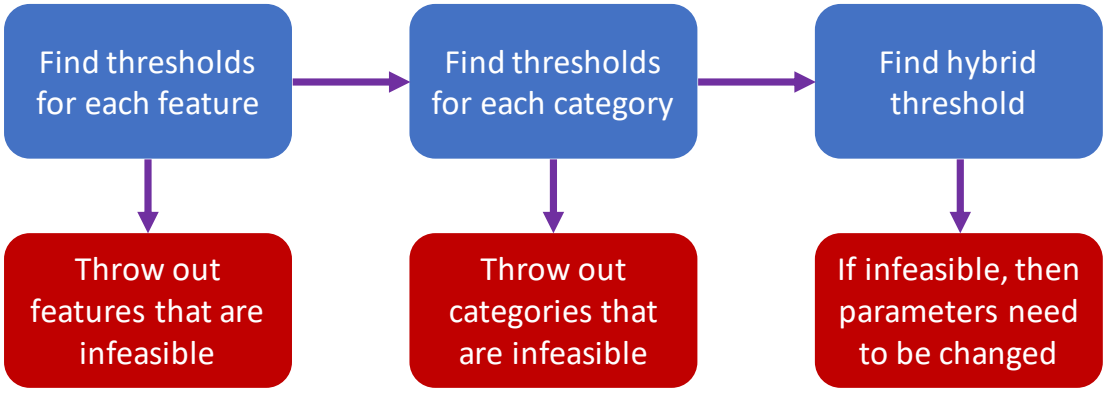

Figure 4-1: **Workflow for our data-driven survival outcome definition.**

Defining the thresholds this way assumes that the outcomes for each category should occur around the same time. This might seem counter-intuitive as some non-motor symptoms, such as sleep problems, are known to occur earlier than motor

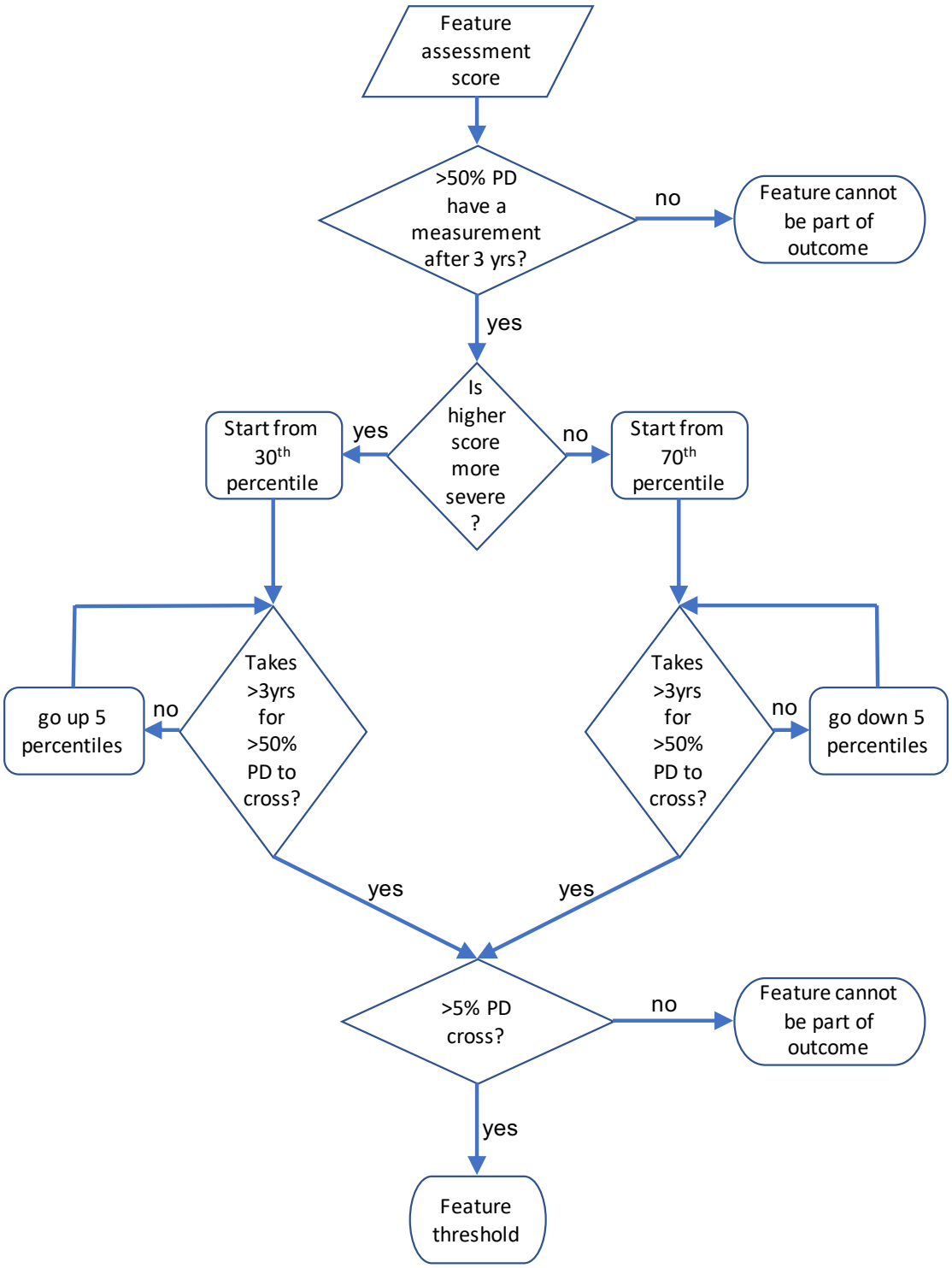

Figure 4-2: **Workflow for selecting the threshold for a single feature or category.**

symptoms. However, our outcomes capture different degrees of severity for each category, and because severity cannot be compared across the different categories anyways, this does not pose an issue. An alternative is to manually set the threshold for each assessment based on literature review. Some assessments have known cutoffs. For example, a score of 25 or below on MoCA indicates mild cognitive impairment [55]. However, we found that a cutoff of 25 would result in a much earlier and less clinically useful outcome. In particular, 22.5% of PD patients are already past that threshold at baseline. Furthermore, it might not capture some aspects of cognitive function, for which other assessments are tailored to measure. These other assessments have no standard thresholds set yet. Therefore, using a data-driven approach would result in a more uniform set of thresholds that are also more tailored for downstream analyses.

To account for noisy observations, we require 2-visit persistence for when a feature crosses a threshold, i.e. a feature must pass a threshold and stay across the threshold the next time it is measured. The first of the two visits is the observation time. In addition, because Parkinson's disease is unlikely to improve over time without treatment, a feature will remain across its threshold once it has crossed, regardless of whether future observations are actually across the threshold.

For motor features, using the totals from MDS-UPDRS part II and III did not satisfy the above requirements. We hypothesize the following two reasons might contribute to this phenomenon: 1) When patients start symptomatic motor treatment, their MDS-UPDRS parts II and III scores would improve. 2) As progression is heterogeneous, patients may exhibit different sets of motor symptoms. In the total score, progression on some set of symptoms may be hidden behind noise in questions on other symptoms. To capture these heterogeneous effects, we instead use the subtotals defined in Table 2.3. As treatment usually only reduces some types of motor symptoms, using subtotals would also allow us to detect progression in the unaffected symptoms. We take the maximum score at a single visit if both on and off exams were taken (this is almost always the off score). Because the motor aspect is the primary hallmark that distinguishes Parkinson's disease patients from prodromal and healthy

control cohorts, we also require significant progression in the motor category for the hybrid outcome.

4.1.2 Derived outcome

The thresholds we found are shown in Table 4.1. CSF and imaging features did not satisfy the criteria above, probably because only a couple measurements were taken for each patient. The survival curves for each category are shown in Figure 4-3. The order in which the categories are crossed in the PD cohort is shown in Table 4.2. Four example PD patients are shown in Figure 4-4. The outcome is observed for the first patient, and its time is determined by nonmotor categories. The outcome is also observed for the second patient, and this time the outcome is determined by the motor category. For the third patient, three of the nonmotor categories are observed, but because the motor outcome does not occur, the patient is censored. The last patient is short one category outcome, so that patient is also censored. As such, we can see that both motor and nonmotor categories play an important role in the outcome definition. Lastly, the proportion of patients who are observed and the average time to event in the PD cohort are shown in Table 4.3, and Figure 4-5 shows the distributions of observation times.

At least 3 of the following motor features:	
MDS-UPDRS II daily activites ≥ 8	MDS-UPDRS III face ≥ 6
MDS-UPDRS III right rigidity ≥ 9	MDS-UPDRS III left rigidity ≥ 9
MDS-UPDRS III tremor ≥ 8	MDS-UPDRS III gait ≥ 3
And at least 2 of the following categories:	
Autonomic: SCOPA-AUT ≥ 11	
Sleep: Epworth $= 1$ and/or REM disorder $= 1$	
Psychiatric: at least 1 of the following:	
QUIP (impulsive) ≥ 1	STAI (anxiety) ≥ 42
GDS (depression) $= 1$	
Cognitive: at least 1 of the following:	
HVLT discrim recog ≤ 6	HVLT immed recall ≤ 14
HVLT retent ≤ 0.4	LNS ≤ 7
BJLO ≤ 8	MoCA ≤ 21
Semantic fluency ≤ 29	

Table 4.1: **Parkinson's disease outcome definition**

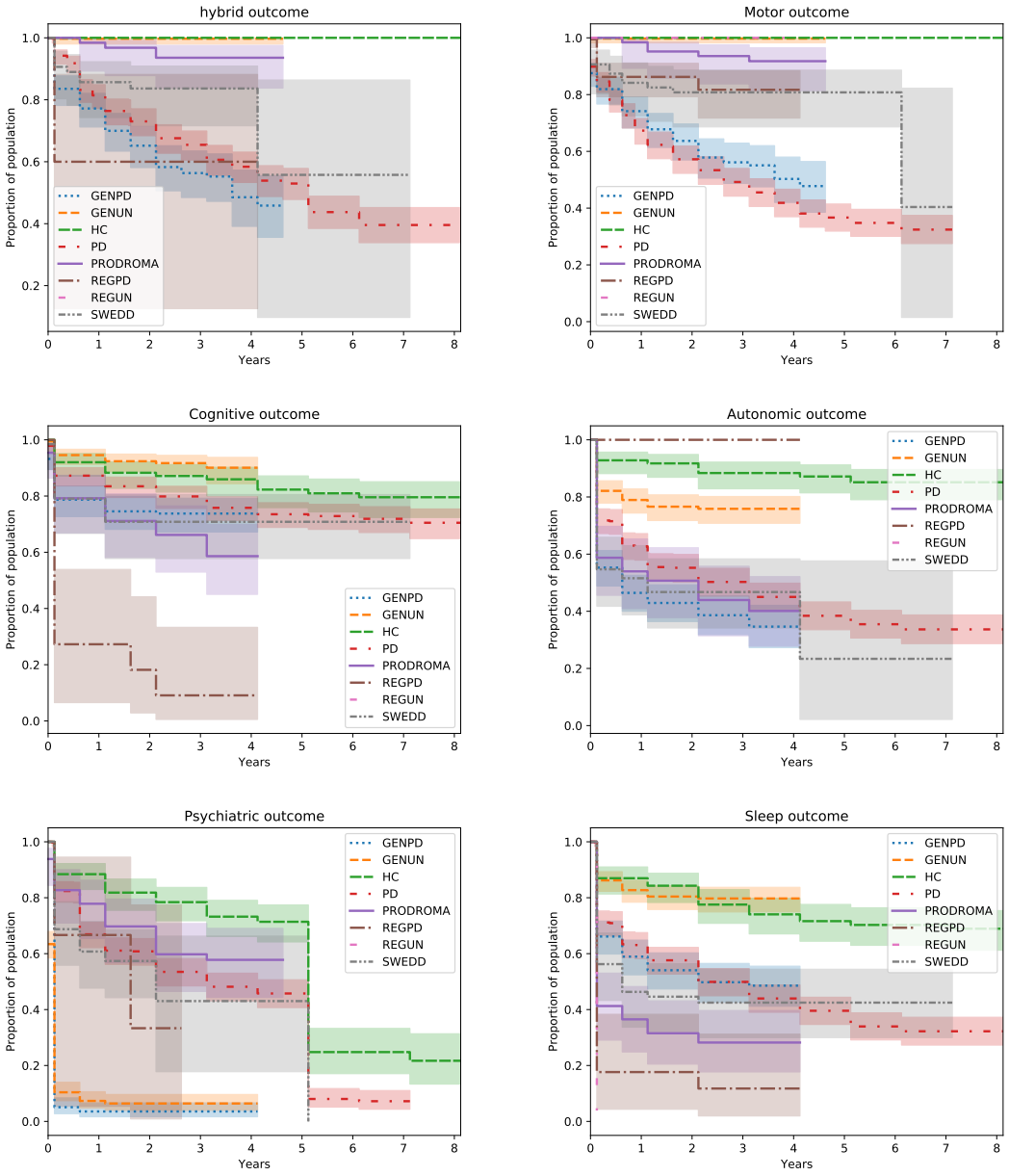

Figure 4-3: **Survival curves** for each outcome and cohort. Shaded regions are confidence intervals generated in Kaplan-Meier plots from lifelines.

Order (first 3)	**Count** (out of 423)
No outcomes observed	28
Psychiatric only	23
Motor only	14
Sleep, Psychiatric only	12
Motor, Psychiatric only	11
{Autonomic, Psychiatric, Sleep}	11
Motor, {Autonomic, Sleep}	11
{Autonomic, Sleep}, Motor	10
Sleep only	8
Sleep, Motor, Psychiatric	7

Table 4.2: **Order of event categories** in **PD** cohort. Ties are denoted with brackets.

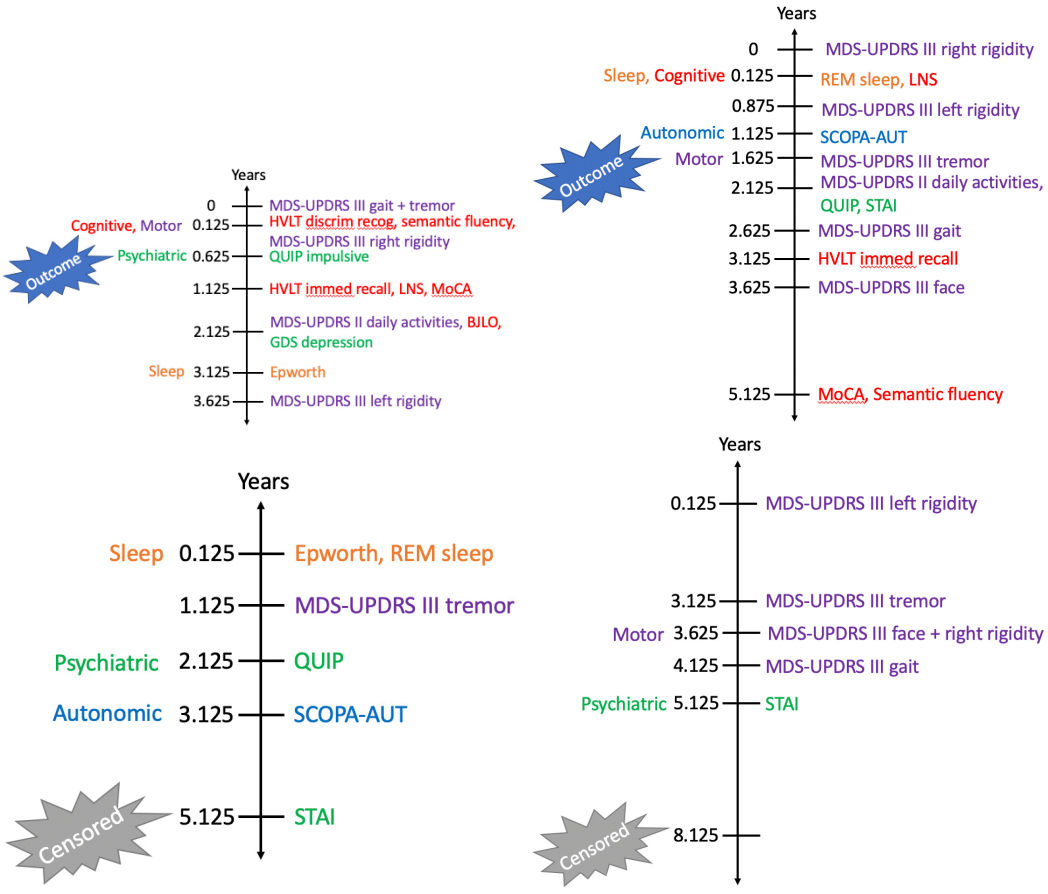

Figure 4-4: **Four example patient timelines**, where the assessment or category name indicates the time at which that threshold was crossed.

Outcome	Prop obs	Time to obs	Time to cens
Autonomic	61.0%	1.28 (1.58)	4.51 (2.13)
Cognitive	25.8%	1.48 (1.68)	4.95 (2.04)
Psychiatric	73.3%	2.31 (2.10)	3.24 (1.96)
Motor	62.4%	1.52 (1.54)	4.70 (1.98)
Sleep	61.9%	1.37 (1.65)	4.44 (2.07)
Hybrid	51.1%	2.24 (1.80)	4.43 (2.01)

Table 4.3: **Outcome statistics** for PD cohort. Average time in years followed by standard deviation in parentheses.

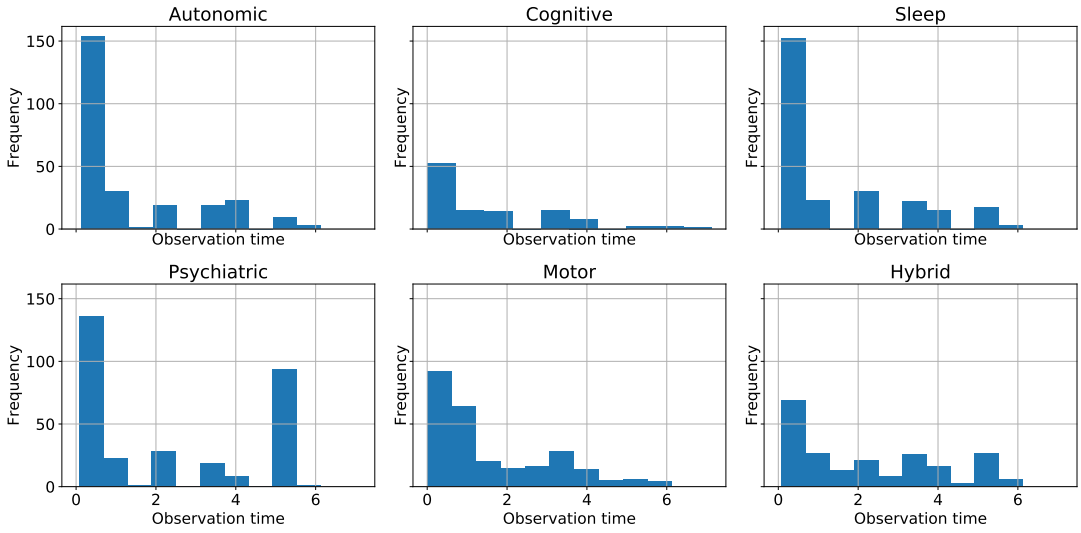

Figure 4-5: **Distribution of outcome observation times** for PD cohort.

4.2 Alternative outcomes

We hope that our outcome definition can aid in future studies of Parkinson's disease and our design criteria opens the door to discussions about other data-driven outcome definitions. Because longitudinal data is available, a series of alternative outcomes can be developed. In this section, we suggest variations for other applications, outcomes that contrast Parkinson's patients and healthy controls, outcomes defined in terms of patient-specific changes, and the consideration of treatments.

4.2.1 Variations for different settings

A common limitation of studies performed using PPMI data is that they cannot be deployed in clinical settings, where some of the assessments used in a model are not taken. The advantage of our data-driven method is that the same criteria can be applied to a different dataset or a different population to create new outcomes. The two parameters in the first criterion (number of years and percentage of cohort) can be tuned to have the desired degree of severity and distribution of survival times for a new application.

Variations on the design criteria may be more useful for certain settings. For example, patients who already have the outcome at baseline would not be useful for a clinical trial, so an example would be to require at most 20% of patients have the outcome at baseline. This is the same design as our first criterion with 50% replaced by 20% and year 3 replaced by year 0. That way, the drug can be tested on a wider patient pool. Another motivation could be maximizing the number of patients who have the outcome within a 3-year observation period. In this case, a criterion that at least 70% are observed within the first 3 years might be added. This would likely have to be used in conjunction with some variant of the first criterion we designed, such as the baseline one, so that the outcomes capture at least some amount of progression.

4.2.2 Outcomes contrasting PD and other cohorts

PPMI has additional cohorts of healthy controls and prodromal patients. As such, it could be advantageous to contrast these patients with the PD cohort when defining the outcome. As we can see in Figure 4-3, the cognitive, autonomic, and sleep outcomes occur earlier in general for the prodromal cohort than the PD cohort. On the surface, this may seem surprising since the PD cohort is much farther along in terms of disease progression. However, this is less surprising when we consider the PPMI eligibility criteria. For one, a prodromal patient must have hyposmia and/or REM sleep behavior disorder at enrollment. As hyposmia is linked to the autonomic outcome and REM sleep behavior disorder is a defining factor of the sleep outcome, prodromal patients are much more disposed to having earlier autonomic and sleep outcomes. Another enrollment criterion is age: at least 30 for the PD cohort and at least 60 for the prodromal cohort. In reality, the average ages upon enrollment for the PD and prodromal cohorts are 61.6 and 68.8, respectively. Since cognitive function is closely tied to age, this might explain why the cognitive outcome tends to occur earlier for the prodromal cohort. Thus, while our outcomes are useful for assessing when a patient will have significant decline in each category of symptoms, potentially affecting choices regarding daily life, we do not claim that our outcomes are solely a reflection of Parkinson's disease progression. Having PD, prodromal, and healthy control cohorts for contrast provides the opportunity to develop methodologies for disentangling the effect of aging and Parkinson's disease. If other datasets have late-stage and early-stage PD patients, contrasting them could also be informative.

4.2.3 Outcomes reflecting relative change

An alternative approach to defining the outcomes would be as changes relative to the patient-specific baseline. There are two potential advantages: 1) This outcome might be more closely related to previous work predicting rates of change. 2) A relative outcome could also be more meaningful since patients who start with more severe symptoms are not already past the outcome. For patients who start with more

moderate symptoms, an intermediate outcome that is closer to their starting point might represent some amount of progression for them. This relative outcome would be observed, while the original outcome might be censored for them.

On the other hand, there are also some drawbacks: 1) If the thresholds are defined in an additive fashion, the assumption that assessments scale linearly with severity is implied. 2) An event no longer represents a uniform amount of severity in one category across patients. 3) For binary features that start at 1, the outcome can never occur. This would counter advantage 2 above, as the amount of censoring in the population could actually increase. This might pose a problem for sleep and psychiatric outcomes, where some of the assessments are binary.

4.2.4 Treatment effect in outcomes

Lastly, we note that our outcome measures do not account for treatment. A patient who reaches the outcome while being treated is likely to have a more severe disease, but the effect on his or her life could be roughly the same as a patient who reached the outcome while untreated. Of course, this does not account for how taking medication likely results in more fluctuations in patient state. For MDS-UPDRS in particular, patients who are on medication sometimes take 2 exams at each visit: one after abstaining from medication the previous night (off exam) and another 1 hour after taking the medication (on exam). We take the maximum subtotal from any exam taken at a single visit. Thus, the motor outcome is primarily driven by "off" and untreated exams. Handling exams that are taken on or off treatment differently may result in outcomes that are more in tune with treatment.

If one were interested in using survival analysis to study treatment effect, the naïve approach would be to first apply the set of definitions or design criteria separately to exams taken while treated and untreated and then taking the difference between the time to untreated and treated outcomes. However, the amount of treatment effect is affected by how severe the patient is when treatment is initiated and how long the patient is on treatment, among other factors. If the outcome time is defined as time since treatment initiation, disease severity at treatment initiation is a confounding

factor for treatment effect. If we only use untreated exams to define the outcome and censor at treatment initiation as suggested by Hernan et al [24], the proportion of censored patients would be much higher. Perhaps some of the approaches in Mackenzie et al [42], such as inverse propensity weighting, could be used once suitable outcome definitions are found.

Chapter 5

Survival analysis

Now that we have a set of outcomes that represent progression events, our next step is to predict time to progression from baseline. There are many ways such a model would contribute to the clinical field. One is identifying features that are important predictors of fast or slow progression. Another is giving patient-specific predictions of time to outcome. Both can help give patients an idea of whether they are slow or fast progressors and if and when they will experience more severe symptoms of Parkinson's. In particular, because our outcomes capture various aspects of PD, predicting them can help better understand the heterogeneity of Parkinson's disease. Lastly, by predicting which patients will have earlier outcomes, we can identify high-risk patients who might benefit from treatment or clinical trial enrollment. In this chapter, we describe two common types of survival models, several sets of covariates we constructed, our inclusion-exclusion criteria, two evaluation metrics, and our results.

5.1 Survival analysis models

In survival analysis, there are 2 functions that are modeled: the survival and hazard functions. The **survival function** maps time t to the probability that the event time will be larger than t, i.e.

$$S(t) = Pr(T > t) \tag{5.1}$$

The **hazard function** measures the event rate at time t given that the event time is larger than t, i.e.

$$h(t) = \lim_{\delta t \to 0} \frac{Pr(t \leq T \leq t + \delta t | T > t)}{\delta t} \tag{5.2}$$

The two are related by the following differential equation:

$$h(t) = -\frac{S'(t)}{S(t)} \tag{5.3}$$

, which solves to

$$S(t) = \exp\left(-\int_0^t h(z)\, dz\right) \tag{5.4}$$

Thus, a larger hazard means the event is more likely to occur.

Survival analysis models differ by how the hazard function is modeled using the baseline covariates. For **Cox proportional hazards**,

$$h(t|x) = \beta_0(t) \exp\left(\sum_{i=1}^{n} \beta_i (x_i - \bar{x}_i)\right) \tag{5.5}$$

This means a 1 unit increase in a covariate x_i translates to the hazard increasing by a factor of $\exp(\beta_i)$. Computing the exact likelihood when there are ties and right censoring is computationally expensive. A common method for approximating the likelihood is the Efron approximation [14].

For the **Weibull accelerated failure time** model,

$$h(t|x) = \left(\frac{t}{\lambda(x)}\right)^{\rho} \tag{5.6}$$

where

$$\lambda(x) = \exp\left(\beta_0 + \sum_{i=1}^{n} \beta_i x_i\right) \tag{5.7}$$

In this case, a larger coefficient β_i actually translates to a lower hazard. Both models are implemented using lifelines [9].

5.2 Covariate sets

For each category, we start by predicting the outcome using only the assessment totals or subtotals that are part of the outcome. These features comprise the baseline covariate sets. Then, we add 3 standard demographic features: age, gender, sense of smell, and disease duration at enrollment to form the standard covariate set. Next, we include treatments in the covariate sets. Cognitive enhancers are rarely used, and PPMI requires that patients are not taking motor medications at baseline. Therefore, no treatment covariates are included in the motor and cognitive covariate sets. Subsequently, we add imaging and CSF biomarkers and then hand-selected relevant features. Following that, we replace the assessment totals and subtotals with the individual questions. We add the genetic risk score from Nalls et al [54] last because it imposes the most severe restriction on sample size. These covariate sets are specified in Tables 5.1 and 5.2. Not shown are two alternate baselines for the motor outcome: using only the MDS-UPDRS part II and III total (MDS) and only the tremor versus postural instability gait dominant indicator (TD/PIGD). These are the simplest ways to obtain a ranking. For the hybrid outcome, all the features under general and each individual category were used in the models.

Only odd BJLO questions are taken at baseline (even and odd questions are used at alternating visits). For questions related to sexual dysfunction in SCOPA-AUT, 0 was imputed for the opposite sex. Use of non-PD related medications (question 26d in SCOPA-AUT) was removed because it is too vague. The following features were removed because they showed no variance across some of the training folds:

- a psychiatric indicator taken during the physical exam (no variance in the more restrictive hybrid outcome patient set; kept for psychiatric models)
- 2 QUIP questions: inability to control gambling and buying
- 3 additional QUIP questions removed only from the hybrid set: too much sex, too much time spent driving or walking with no goal in mind, and too much gambling

- MoCA questions: name city, place, day, month, year, lion, and camel
- LNS questions 1a, 5a, 5b, 5c, 6a, 6b, 6c, 7a, 7b, and 7c (first four omitted due to no variance and last five because they are missing for too many patients)
- indicators for narcolepsy, epilepsy, and inflammatory disease of brain in REM sleep questionnaire
- freezing of gait question from MDS-UPDRS part III

Features	**BL**	**Std**	**Std+Trt**	**Std+Trt+Img+CSF**
General features				
Age, Male, UPSIT, disease duration		4	4	4
Genetic risk score				
CSF biomarkers				7
imaging biomarkers				8
Motor features				
MDS-UPDRS II + III subtotals	6	6	6	6
Hoehn & Yahr				
MDS-UPDRS II + III questions				
TD / PIGD				
Cognitive features				
Assessment totals in outcome	7	7	7	7
MDS-UPDRS I cognitive				
BJLO questions				
MoCA questions				
LNS questions				
Semantic fluency questions				
HVLT statistics				
Autonomic features				
SCOPA-AUT total	1	1	1	1
Digestive aid			1	1
Medical history gastrointestinal				
Blood pressures				
BMI				
MDS-UPDRS 1.9-1.13				
SCOPA-AUT questions				
Psychiatric features				
GDS, QUIP, + STAI totals	3	3	3	3
Antidepressants + anxiolytics			2	2
Current + history psychiatric				
White race				
MDS-UPDRS 1.2-1.5				
GDS questions				
QUIP questions				
STAI questions				
Sleep features				
Epworth + REM sleep totals	2	2	2	2
Sleep aid			1	1
MDs-UPDRS 1.7-1.8				
Epworth questions				
REM sleep questions				

Table 5.1: **Four sets of covariates** for survival models. BL: baseline. Std: standard. Trt: treatment. Img: imaging. Entry indicates the number of features included in the covariate set (empty is 0). The other three covariate sets are in the next table.

Features	**Std+Trt+Img +CSF+Exp**	**Qst+Trt+Img +CSF+Exp**	**Qst+Trt+Img +CSF+Exp+Gen**
General features			
Age, Male, UPSIT, disease duration	4	4	4
Genetic risk score			1
CSF biomarkers	7	7	7
Imaging biomarkers	8	8	8
Motor features			
MDS-UPDRS II + III subtotals	6		
Hoehn & Yahr	1	1	1
MDS-UPDRS II + III questions			45
TD / PIGD	2	2	2
Cognitive features			
Assessment totals in outcome	7		
MDS-UPDRS I cognitive	1	1	1
BJLO questions		15	15
MoCA questions		19	19
LNS questions		11	11
Semantic fluency questions		3	3
HVLT statistics		7	7
Autonomic features			
SCOPA-AUT total	1		
Digestive aid	1	1	1
Medical history gastrointestinal	1	1	1
Blood pressures	4	4	4
BMI	1	1	1
MDS-UPDRS 1.9-1.13	5	5	5
SCOPA-AUT questions		30	30
Psychiatric features			
GDS, QUIP, + STAI totals	3		
Antidepressants + anxiolytics	2	2	2
Current + history psychiatric	2	2	2
White race	1	1	1
MDS-UPDRS 1.2-1.5	4	4	4
GDS questions		15	15
QUIP questions		11	11
STAI questions		40	40
Sleep features			
Epworth + REM sleep totals	2		
Sleep aid	1	1	1
MDs-UPDRS 1.7-1.8	2	2	2
Epworth questions		8	8
REM sleep questions		18	18

Table 5.2: **Three other sets of covariates** for survival models. Exp: expanded. Qst: questions. Gen: genetic. See previous table for the other four covariate sets and description.

5.3 Inclusion-exclusion criteria

As PPMI enrollment criteria resembles clinical trial enrollment criteria, we did not have to add additional criteria regarding patient eligibility. For modeling purposes, we imposed two additional criteria:

1. No missing baseline features

2. Observation or censoring time is after time 0.

In particular, to have a unified test set across all the covariate sets, the test set is 20% of the cohort that satisfies the criteria above when considering the union of baseline features used in any model. Note that the test set is different for each outcome. Models with different covariate sets may be trained and validated on different sets of samples. Any sample that is not part of the test set and satisfies the two criteria above for the set of covariates used by a model are part of the training and validation set. Training and validation are split 75/25 in a 4-fold cross-validation such that each patient occurs in the validation set of exactly one fold.

One concern with this method is that patients with missing baseline features might differ from patients with no missing features. To examine this hypothesis, we compared the included and excluded cohorts in terms of the observation and censoring statistics in Table 5.3, distributions of observation times in Figure 5-1 and distributions of the standard set of baseline covariates in Figures 5-2 to 5-7. The distributions in the figures are similar across the included and excluded cohorts. The excluded cohorts seem to have a bit more censoring for some of the outcomes. This may be because patients who are missing baseline features may also be missing later measurements, and the outcome is less easily detected in patients with fewer measurements. The distributions of baseline features seem similar across the included and excluded cohorts for all outcomes. Overall, it seems like little bias is induced by applying these inclusion-exclusion criteria to the PD cohort. The number of patients that were available for each outcome model are shown in Figure 5-8.

Outcome	# pat.	Prop obs	Time to obs	Time to cens
Autonomic	291/127	64.3%/55.9%	1.22(1.57)/1.46(1.61)	4.59(1.96)/4.77(2.10)
Cognitive	322/75	25.2%/25.3%	1.73(1.73)/1.13(1.45)	5.26(1.70)/5.14(1.71)
Psychiatric	290/125	73.8%/76.8%	2.17(2.07)/2.62(2.13)	3.36(1.86)/3.81(1.65)
Motor	311/69	60.1%/49.3%	1.84(1.50)/1.80(1.72)	4.78(1.88)/4.39(2.30)
Sleep	331/85	64.7%/56.5%	1.40(1.68)/1.26(1.53)	4.80(1.76)/4.17(2.19)
Hybrid	268/148	58.2%/40.5%	2.13(1.74)/2.53(1.94)	4.65(1.82)/4.50(1.93)

Table 5.3: **Statistics for patients included and excluded due to first criterion** using union of all covariate sets. Only patients who satisfy second criterion are part of these statistics. Included statistics precede the slash (/). Excluded statistics follow it. # pat. stands for number of patients. Average time in years followed by standard deviation in parentheses.

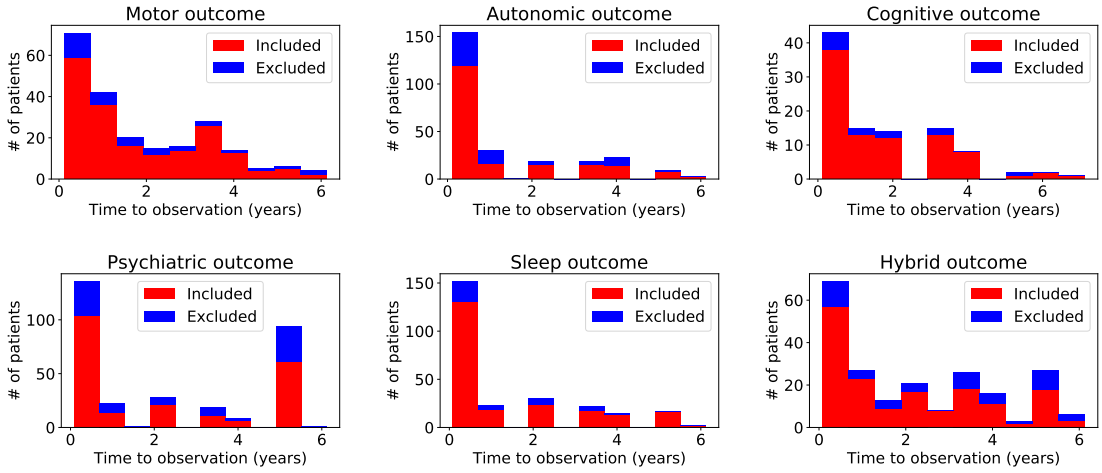

Figure 5-1: **Distribution of outcome observation times in included and excluded** cohorts.

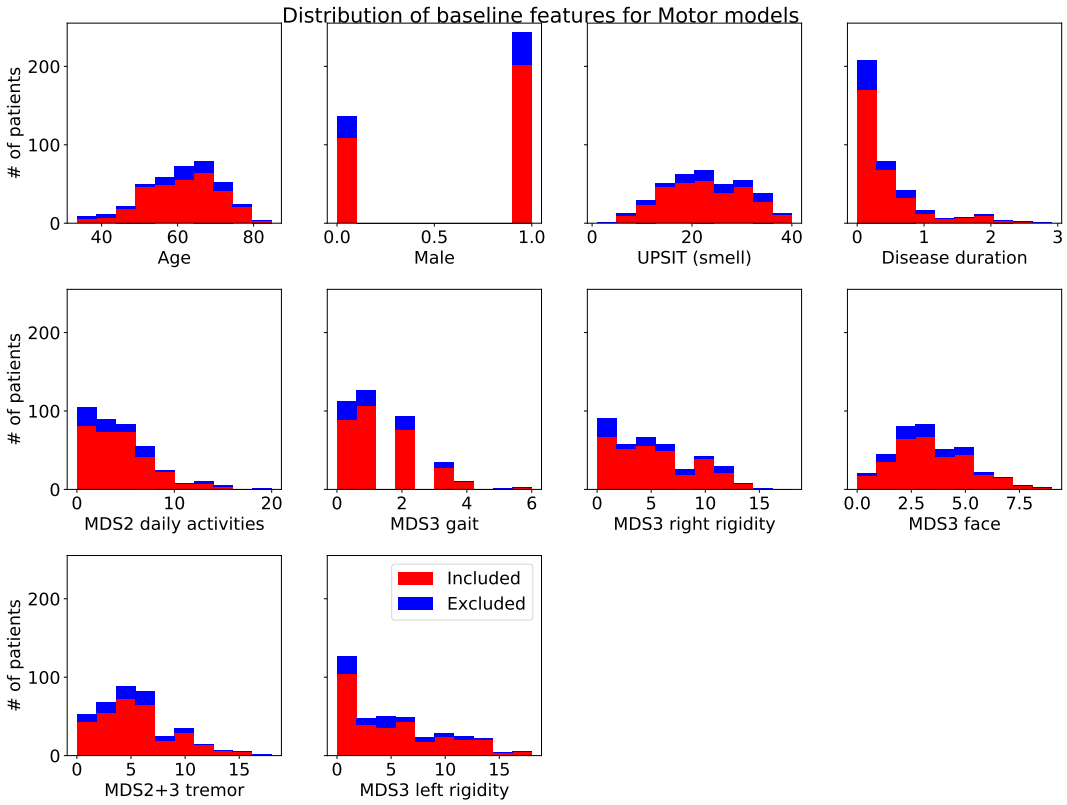

Figure 5-2: **Distribution of baseline features for included and excluded** cohorts for **motor** outcome.

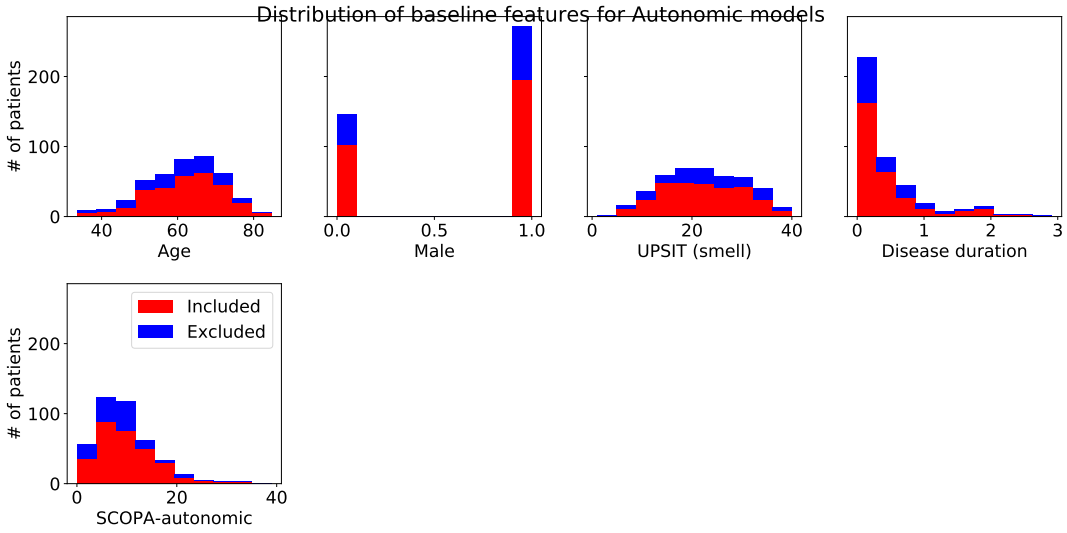

Figure 5-3: **Distribution of baseline features for included and excluded** cohorts for **autonomic** outcome.

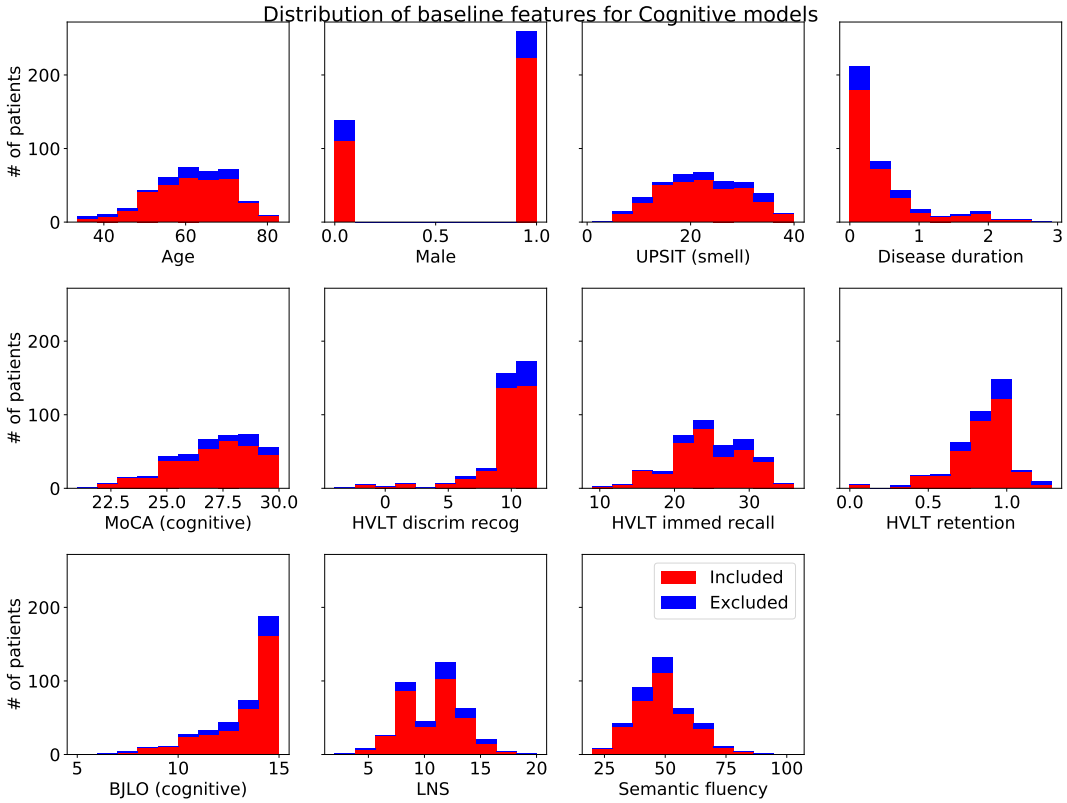

Figure 5-4: **Distribution of baseline features for included and excluded** cohorts for **cognitive** outcome.

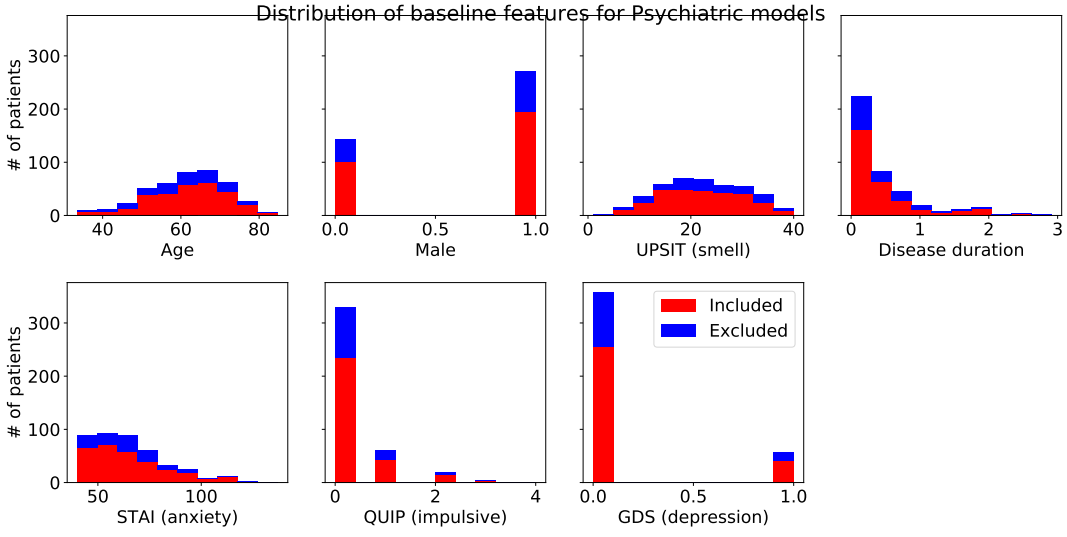

Figure 5-5: **Distribution of baseline features for included and excluded** cohorts for **psychiatric** outcome.

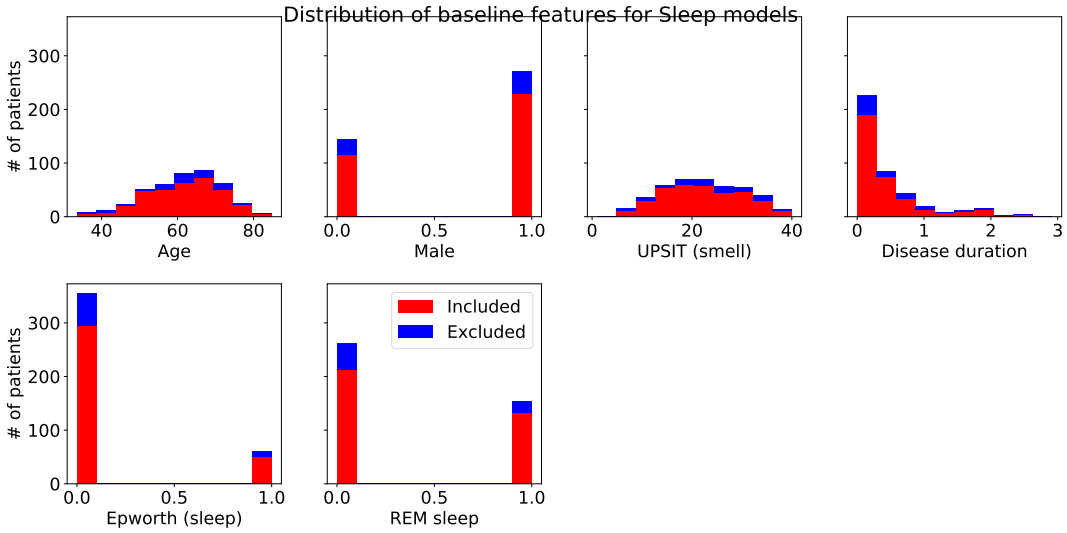

Figure 5-6: **Distribution of baseline features for included and excluded** cohorts for **sleep** outcome.

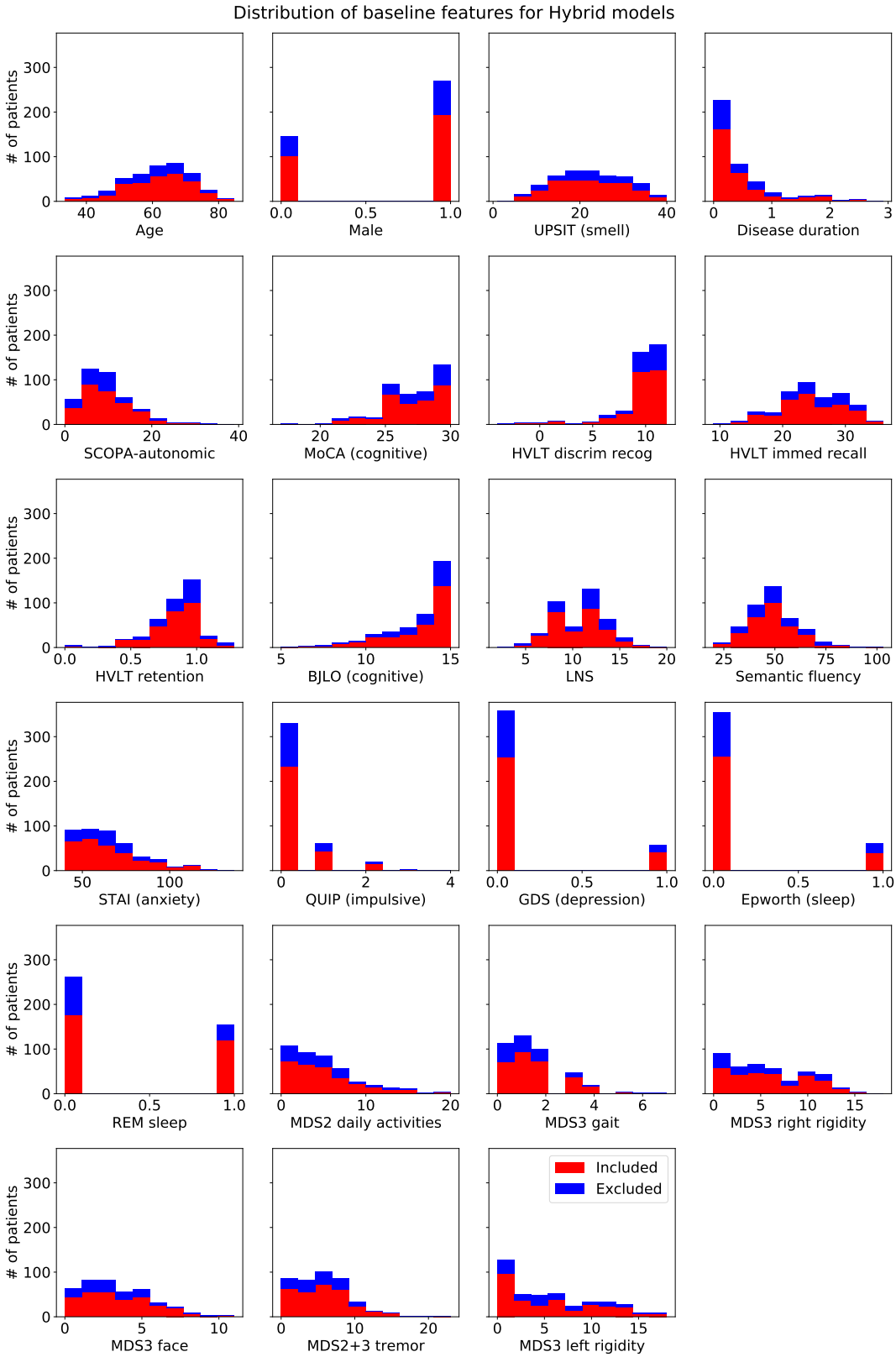

Figure 5-7: **Distribution of baseline features for included and excluded** cohorts for **hybrid** outcome.

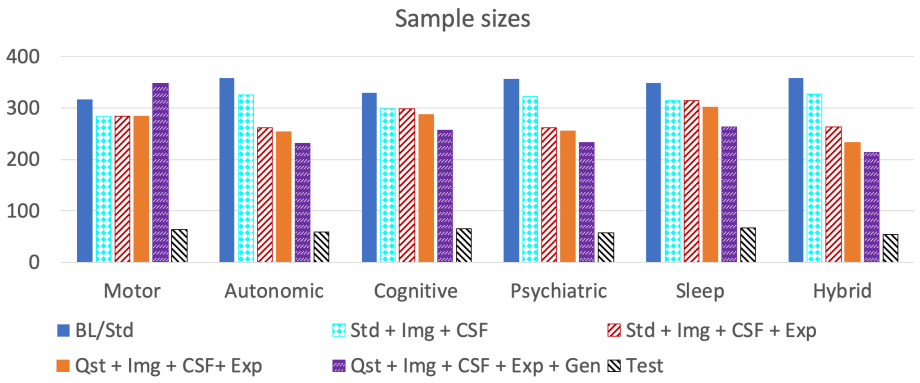

Figure 5-8: **Sample sizes for survival analysis models**

5.4 Evaluation metrics

Survival analysis models are primarily evaluated using two metrics: concordance index and mean absolute error. For both, predictions that exceeded the maximum time to censoring or observation in the training set were truncated to be at that maximum time. The concordance index (CI) measures the proportion of event pairs that are ordered correctly:

$$CI = \frac{1}{|\mathcal{E}_{ij}|} \sum_{\mathcal{E}_{ij}} 1\left\{\widehat{T}_j > \widehat{T}_i\right\} \tag{5.8}$$

, where $\mathcal{E}_{ij}$ is the set of pairs of events such that event i is observed at time T_i and event j is censored or observed after T_i at T_j. $\widehat{T}_i$ and $\widehat{T}_j$ are the predicted event times. When calculating mean absolute error (MAE), we account for censoring by taking a one-sided error for censored patients, i.e. any prediction after the censoring time is not penalized.

5.4.1 Cross-validation and regularization

The validation set in each of the 4 folds was used for selecting the best regularization settings. For Cox models, we tuned the amount of regularization (0, 0.01, 0.1, 0.5, 1, 1.5, 2, 3, 4, 5, 7, 10, 15, 20, 25, 35, 50, 65, 80, 100, 120, 150, 200, 250, 300, 350, 400, 450, 500, 600, 700, 800, 900, 1000, 1100, 1200, 1300, 1500, 1750, 2000, 2250, 2500, 3000, 3500, 4000, 4500, 5000, 6000, 7000, 8000, 9000, 10000, 11000, 12500, and 15000. Some of the larger regularization settings were omitted for smaller covariate sets.). Only L2 regularization was available for Cox models in lifelines. For Weibull models, we tuned the amount of regularization (0, 0.01, 0.05, 0.1, 0.4, 0.7, 1, 1.5, 2, 2.5, 3, 3.5, 4, 4.5, 5, 6, 7, 10, 15, 20, 25, 35, 50, 65, 80, 100) and the ratio between the amount of L1 and L2 regularization (0, 0.5, and 1). For each fold, the regularization setting with the highest validation CI was selected. If two or more settings have the same validation CI, then from among them, the setting with the lowest validation MAE is selected.

The default initialization for the parameters is 0 in lifelines. If lifelines encounters

an error while training, up to 9 other random initializations drawn from a standard Normal distribution were tried. If all 10 initializations fail to converge, the parameter setting is skipped. The goal is to avoid having no convergence in all parameter settings.

First, we examine the regularization plots for the Weibull model to ensure that the regularization settings we tried were sufficient to possibly cover the best setting. In the top row of Figure 5-9, we see that the validation concordance index starts to drop off to the right of the selected setting, indicating that there is too much regularization. The effect of underfitting is much more obvious when looking at MAE. In particular, we note that in folds 2 and 3, the optimal setting selected using validation CI does not align with the optimal setting had validation MAE been used instead. To the left of the selected setting, the U-shape in the top row and the slight increase in the bottom row are a bit odd since they suggest a little bit of regularization performs worse than no regularization. Figure 5-10 shows that using only L2 regularization rather than a mix of L1 and L2 regularization was the best option for the Weibull model for predicting the cognitive outcome. As an aside, the way validation CI is consistently higher and validation MAE is consistently lower when compared with the training metrics in fold 2 is likely because the validation set is easier to predict than the training set and not a sign of any issues with training the models.

Turning to Cox proportional hazards, first looking at the regularization plot in Figure 5-11 for a model with a small number of covariates, we see that a small amount of regularization makes little difference. When a larger amount of regularization is applied, both of the metrics suddenly worsen. In this case, regularization was probably unnecessary. In Figure 5-12, for the hybrid outcome with a large covariate set, we need to use much larger penalizers to reach the underfitting trend on the right. The spikes of poor performance and x's indicating errors during training suggest that there are some cases where the Efron approximation method used by lifelines for computing the likelihood does not work well.

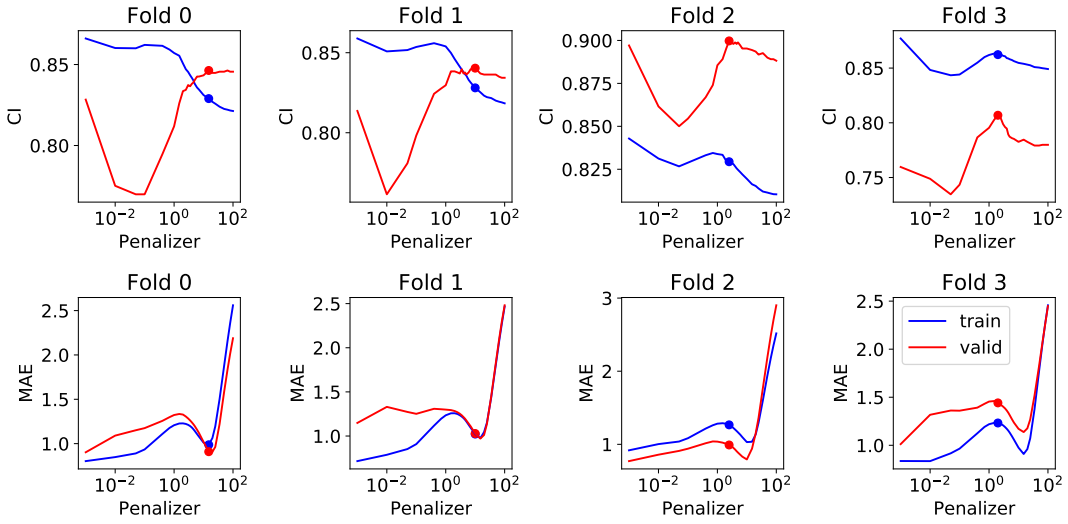

Figure 5-9: Train and validation evaluation metrics from **varying the penalizer in the Weibull** model across the 4 folds. Model shown is for predicting the cognitive outcome from the standard covariate set, which obtained the highest CI. Selected setting is shown with a point. x indicates lifelines encountered an error in that setting and no model was trained.

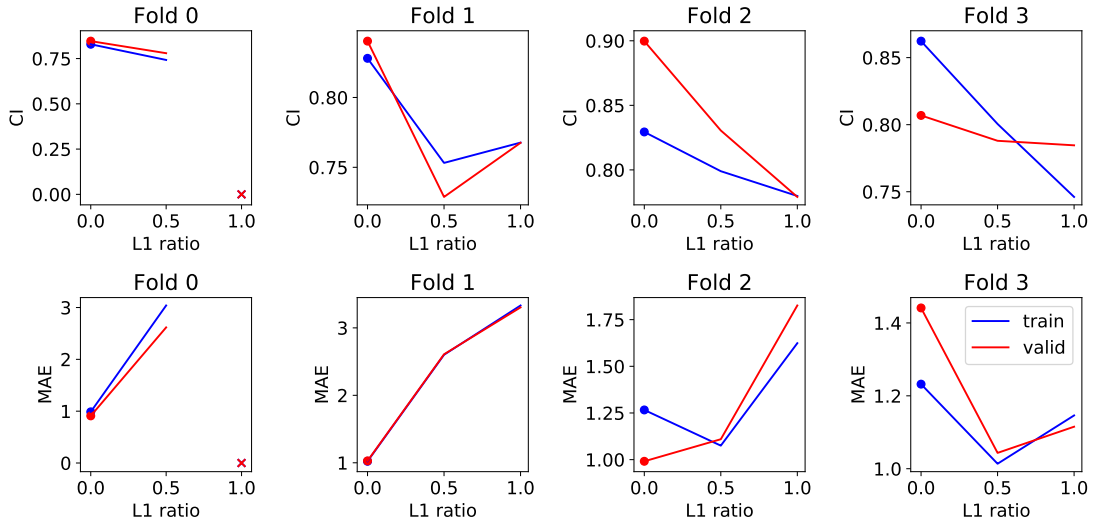

Figure 5-10: Train and validation evaluation metrics from **varying L1 ratio in the Weibull** model across the 4 folds. See previous figure for model and figure description.

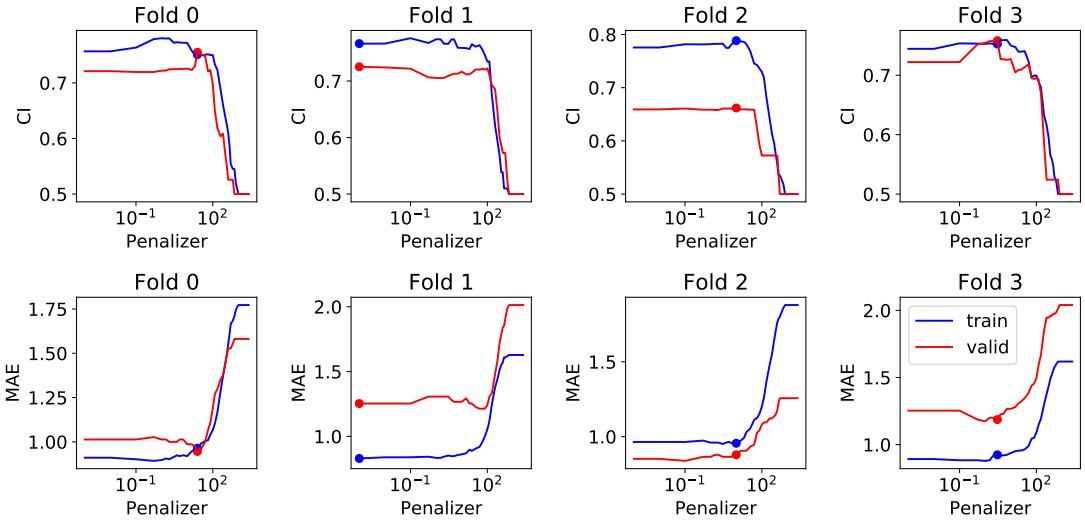

Figure 5-11: Train and validation evaluation metrics from **varying the penalizer in Cox** proportional hazards across the 4 folds. Model shown is for predicting the cognitive outcome from the standard + imaging + CSF covariate set, which obtained the lowest MAE. Selected setting is shown with a point.

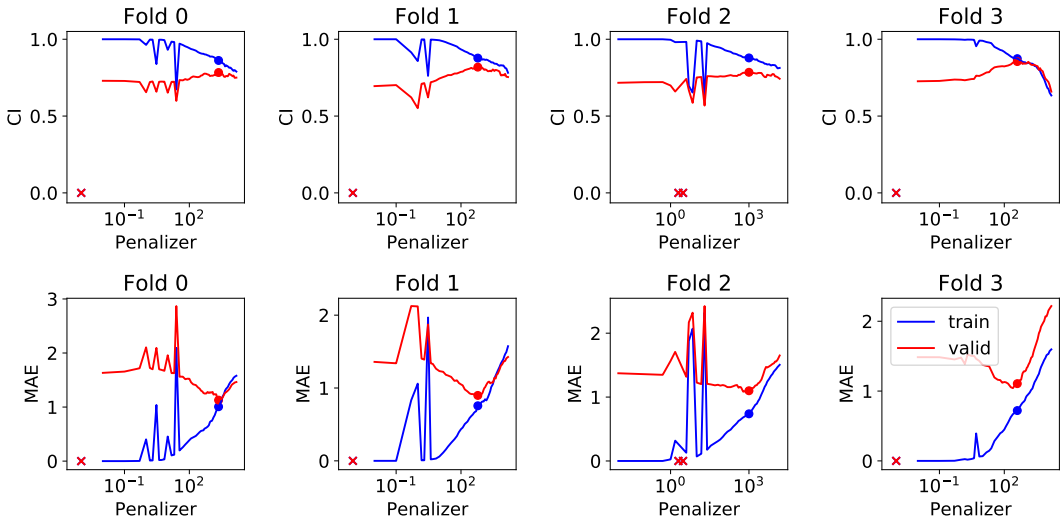

Figure 5-12: Train and validation evaluation metrics from **varying the penalizer in the Weibull** model across the 4 folds. Model shown is for predicting the hybrid outcome from the questions + treatment + imaging + CSF + expanded covariate set. Selected setting is shown with a point. x indicates lifelines encountered an error in that setting and no model was trained.

5.5 Results and discussion

First, we examine how well our models performance using the quantitative evaluation metrics defined above and select the best covariate sets for each outcome. Then, we look at the coefficients to understand what features drive the models. Lastly, we discuss some limitations to our work.

5.5.1 Quantitative evaluation

For each outcome, models using different covariate sets were evaluated on the same test set. Recall that by the definition of CI, randomly ranking the patients would in expectation result in a CI of 0.5. Likewise, predicting a constant time for all patients would also result in a CI of 0.5. We verified this was the case using the lifelines implementation, even when ties were present. As we can see in Figures 5-13 to 5-18, all of the CIs were above 0.5, indicating that the models were at least better than random.

For MAE, two simple baselines were predicting the median among only the observation times and among both the observation and censoring times. For the motor outcome, predicting the median observation time had lower MAE than any of the models, suggesting there is still room for improvement in predicting the motor outcome. The other outcomes have better models.

In general, Weibull models perform better on MAE because the objective of Cox proportional hazards is to get the correct relative ordering, not absolute times. Cognitive was an exception to this pattern because most patients are censored. Because there are fewer observed patients to compare against for ranking, CI was generally much higher for the cognitive outcome. Similarly, because only a one-sided error is taken for censored patients, MAE was also much lower.

In particular, we note that enlarging the covariate set does not necessarily lead to improved performance. Theoretically, with regularization, the coefficients for the added covariates could be set to 0, so model performance should be at least as good as a model with a subset of the covariates. The deterioration we observed may in

part be due to reduced sample size. Statistical power is reduced when too many features are included in the dataset. To avoid drawing incorrect conclusions from the coefficients, examining different covariate sets is important.

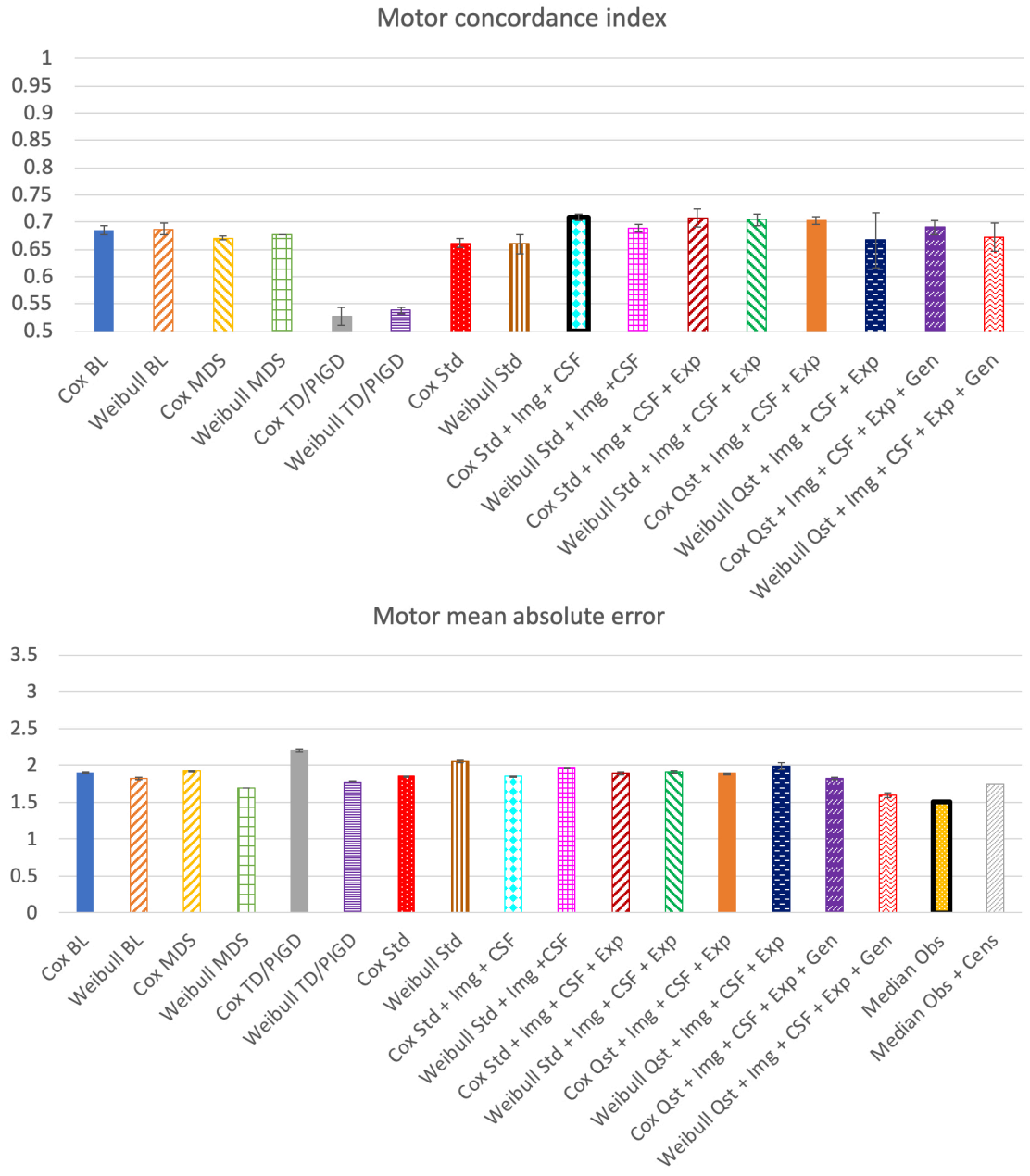

Figure 5-13: **Evaluation of survival analysis models for motor outcome** prediction from baseline features. Error bars are standard deviation from 4-fold cross evaluation. Bars for the best model for each outcome according to each metric have a bolded black border.

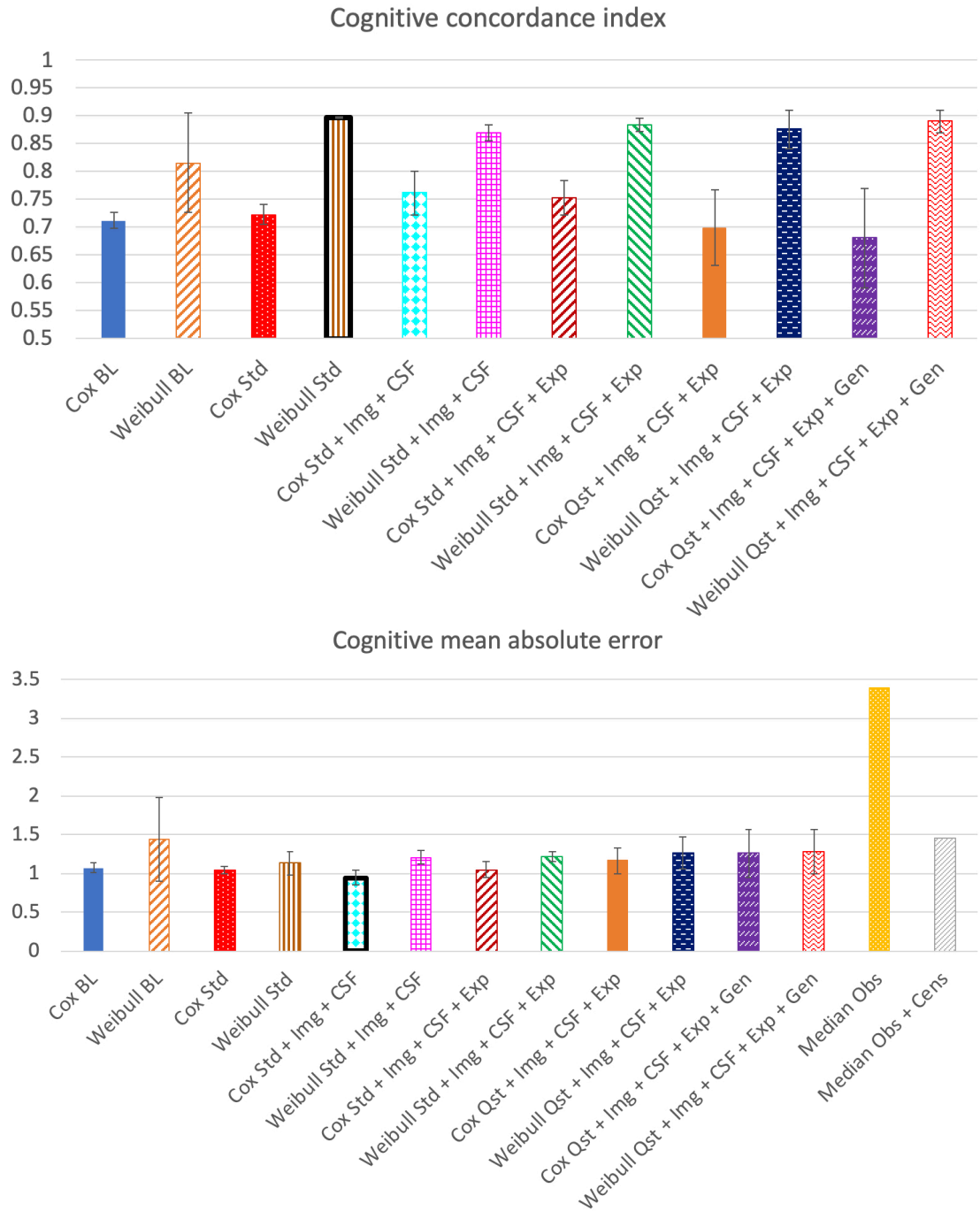

Figure 5-14: **Evaluation of survival analysis models for cognitive outcome** prediction from baseline features. See Figure 5-13 for description.

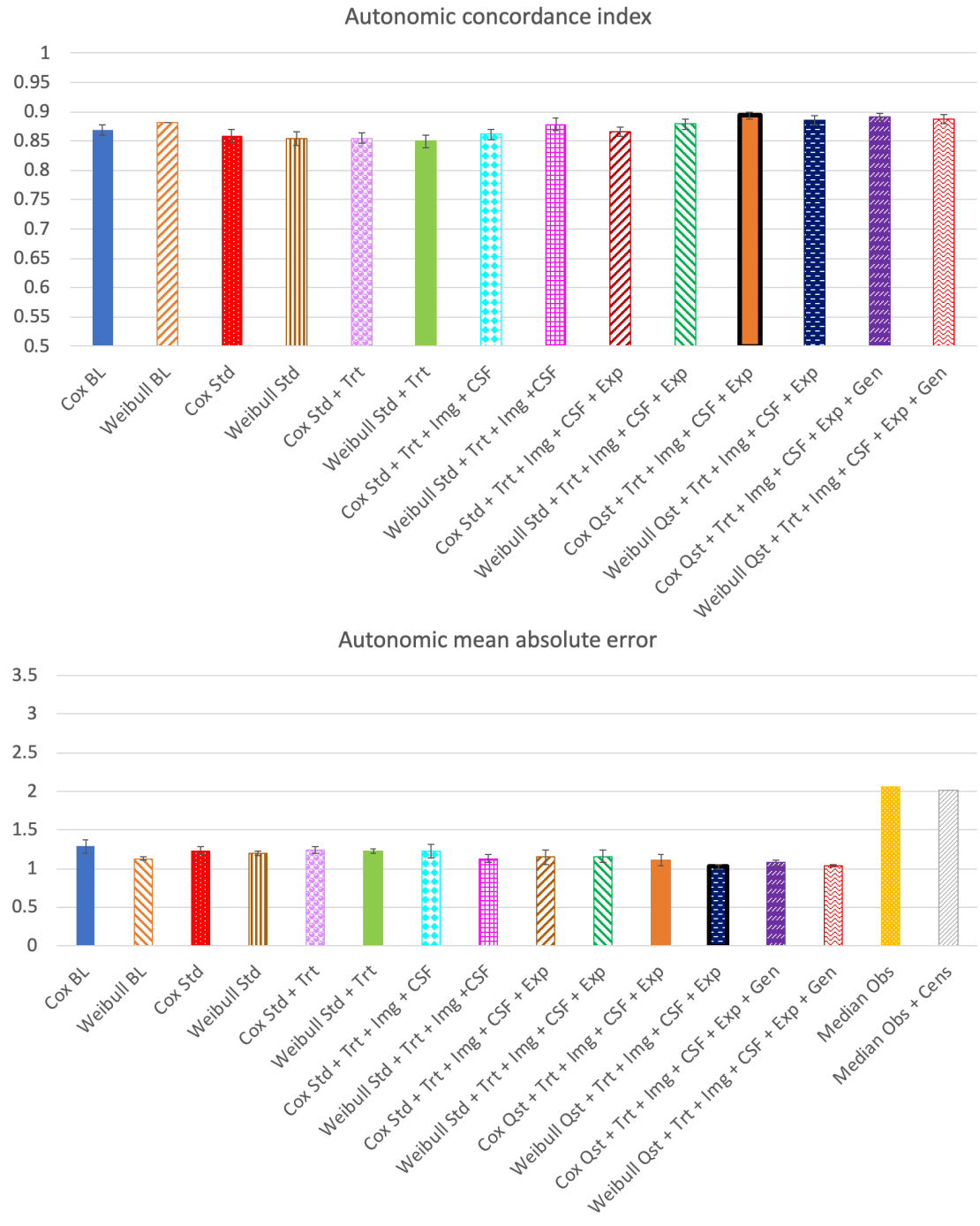

Figure 5-15: **Evaluation of survival analysis models for autonomic outcome** prediction from baseline features. See Figure 5-13 for description.

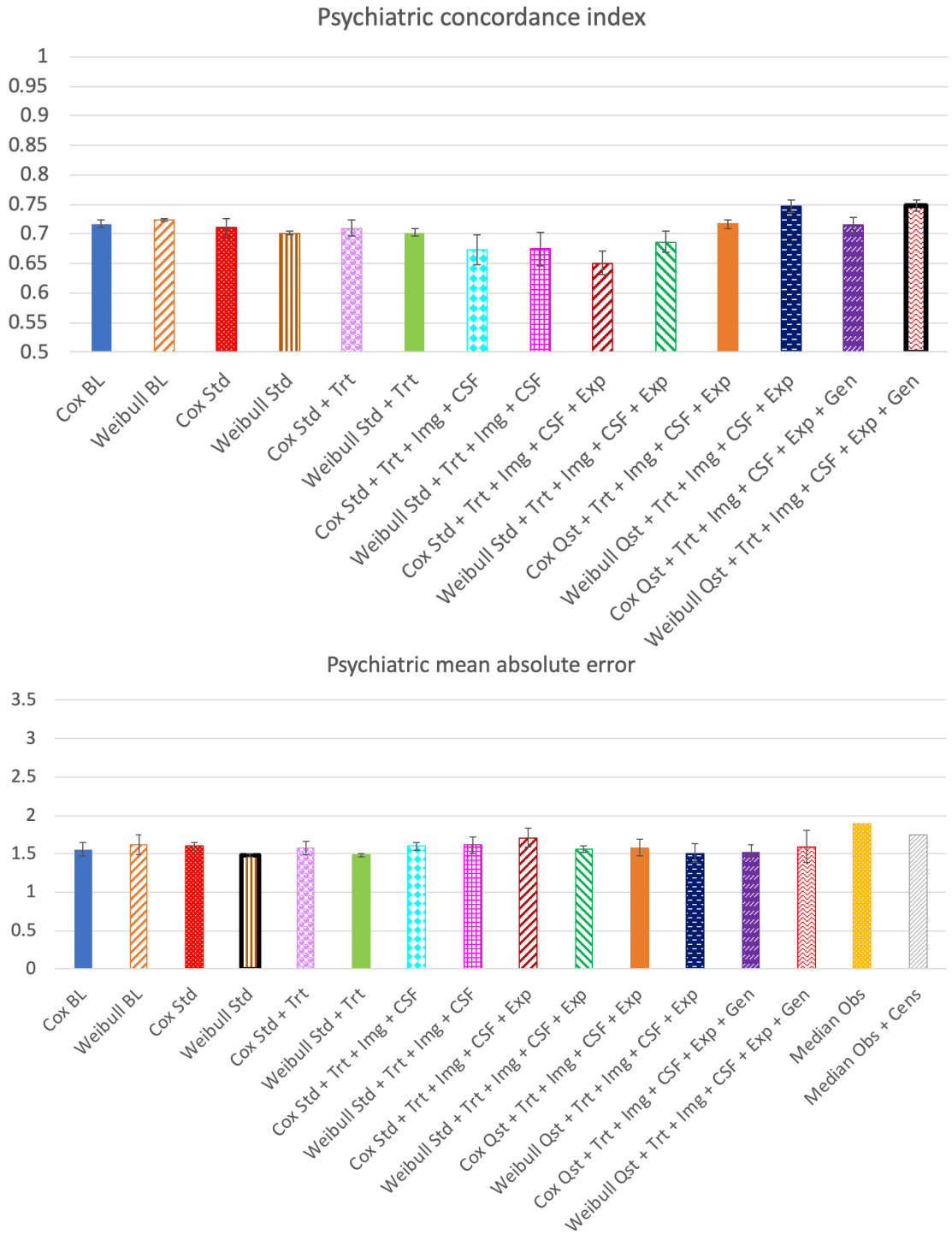

Figure 5-16: **Evaluation of survival analysis models for psychiatric outcome** prediction from baseline features. See Figure 5-13 for description.

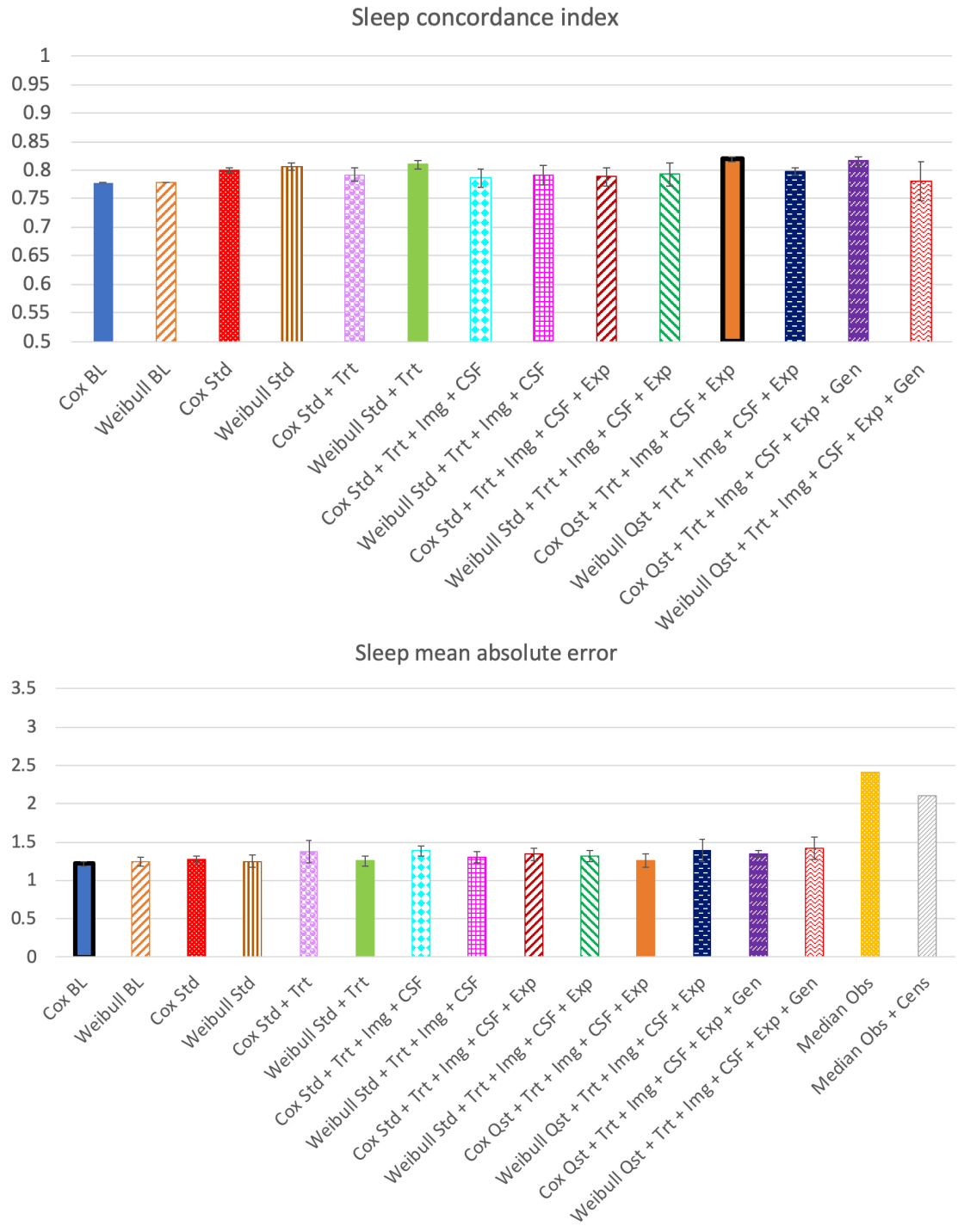

Figure 5-17: **Evaluation of survival analysis models for sleep outcome** prediction from baseline features. See Figure 5-13 for description.

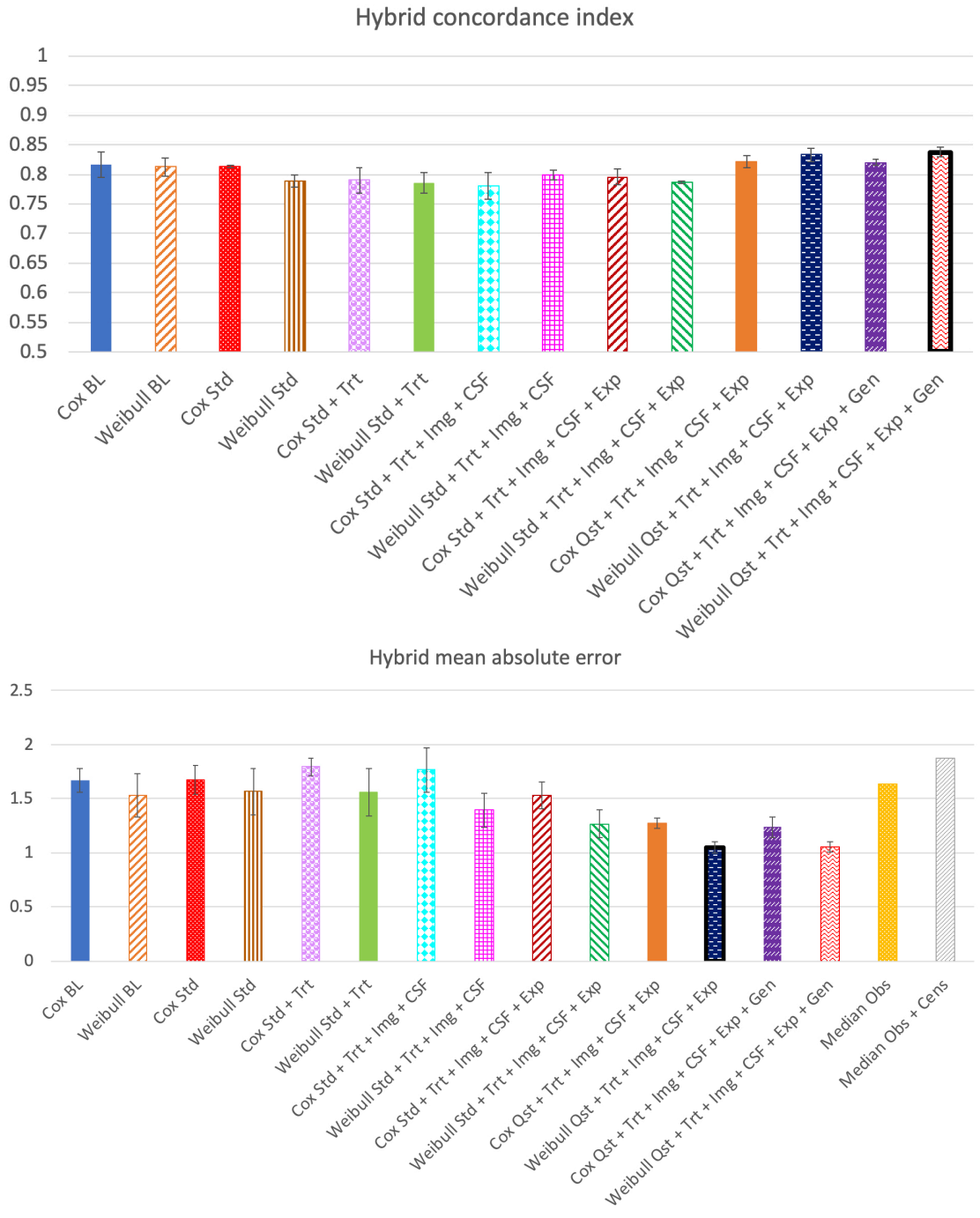

Figure 5-18: **Evaluation of survival analysis models for hybrid outcome** prediction from baseline features. See Figure 5-13 for description.

5.5.2 Qualitative discoveries

Before we start examining the coefficients in detail, recall that the coefficients in Cox and Weibull models are supposed to have opposite signs as an increase in hazard is indicated by a positive coefficient in Cox and a negative coefficient in Weibull. For each category, we show the coefficients from the model with the highest CI and the model with the lowest MAE in the figures and tables in this section. Models that include questions in the covariate set have a large number of features, so only those that have significant coefficients are shown in the tables. By significant, we mean the mean across the 4 folds is at least 1.96 standard deviations away from 0 (i.e. p-value is less than 0.05 in a two-sided hypothesis test).

Feature	**Coefficient**
White	0.1236 (0.0441)
STAI steady	0.0655 (0.0314)
STAI 36 content	0.0622 (0.0313)
STAI 21 pleasant	0.0565 (0.0223)
tTau (log)	0.0558 (0.0241)
STAI 33 secure	0.0332 (0.0169)

Table 5.4: **Coefficients for psychiatric outcome** from Weibull model with question set of covariates. This model had the highest concordance index. Standard deviation across 4 folds shown in parentheses. Of the 95 features, only the 6 shown here had significant coefficients.

In Figure 5-19, we can see that all of the motor subtotals except tremor are predictive of earlier motor outcome. Tremor is not predictive of early or late outcomes. Hyposmia (indicated by a lower score on UPSIT) is also predictive of an earlier outcome. We see large coefficients on some of the imaging and CSF biomarkers, which often occurs in models for other outcomes as well. We hypothesize this may be an artifact of some of the features having right-skewed distributions, as seen in Figures 2-8 and 2-9. After min-max scaling, these features have smaller magnitudes than other features, which might be the cause of larger coefficients. Scaling these features into a Normal distribution may have been a better choice.

Figure 5-20 shows that all of the cognitive assessments contribute to the cognitive outcome prediction, as expected. It also shows that smell (UPSIT) is predictive.

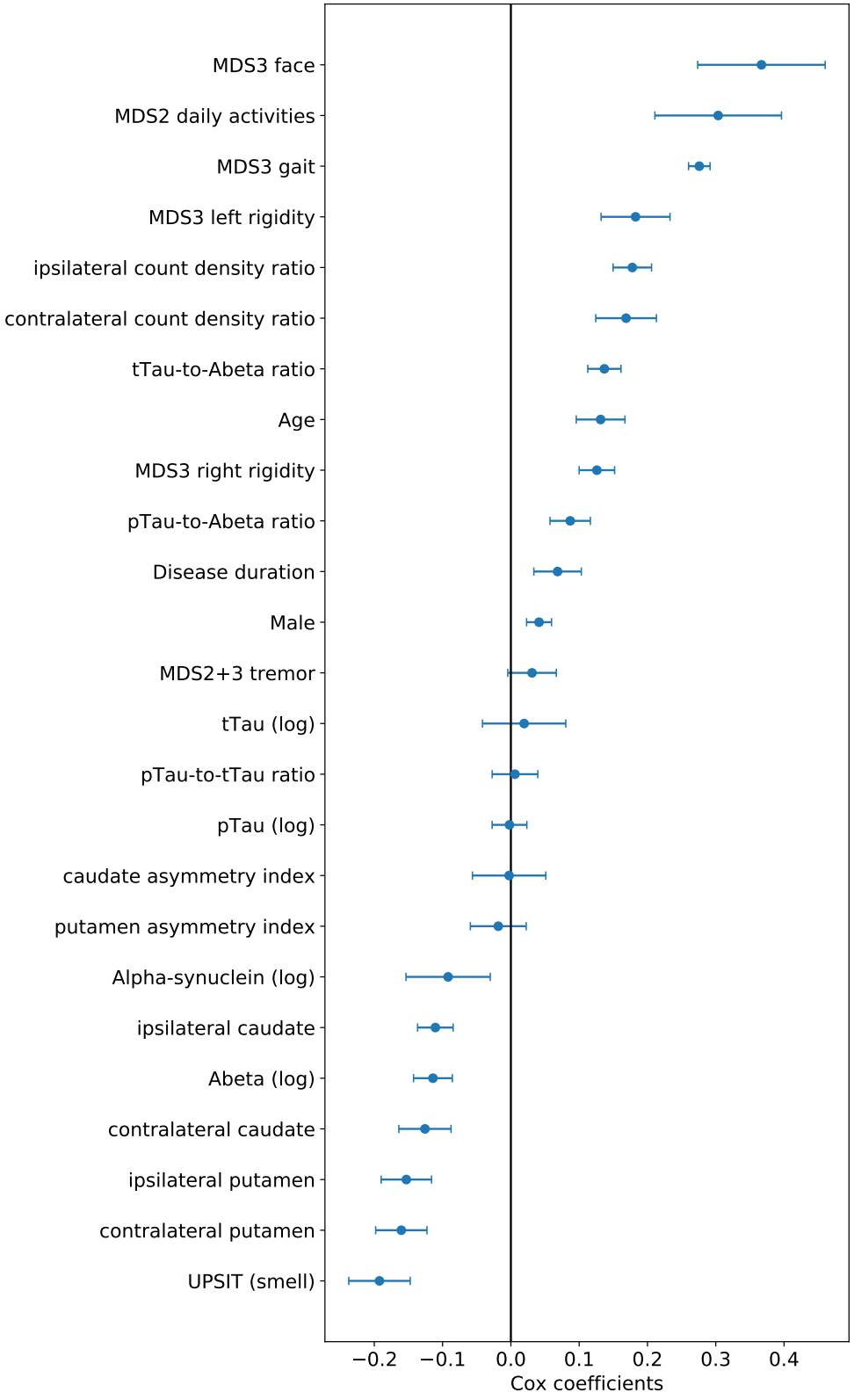

Figure 5-19: **Coefficients for motor outcome** from Cox model with standard + imaging + CSF set of covariates. This model had the highest concordance index. Standard deviation across 4 folds shown.

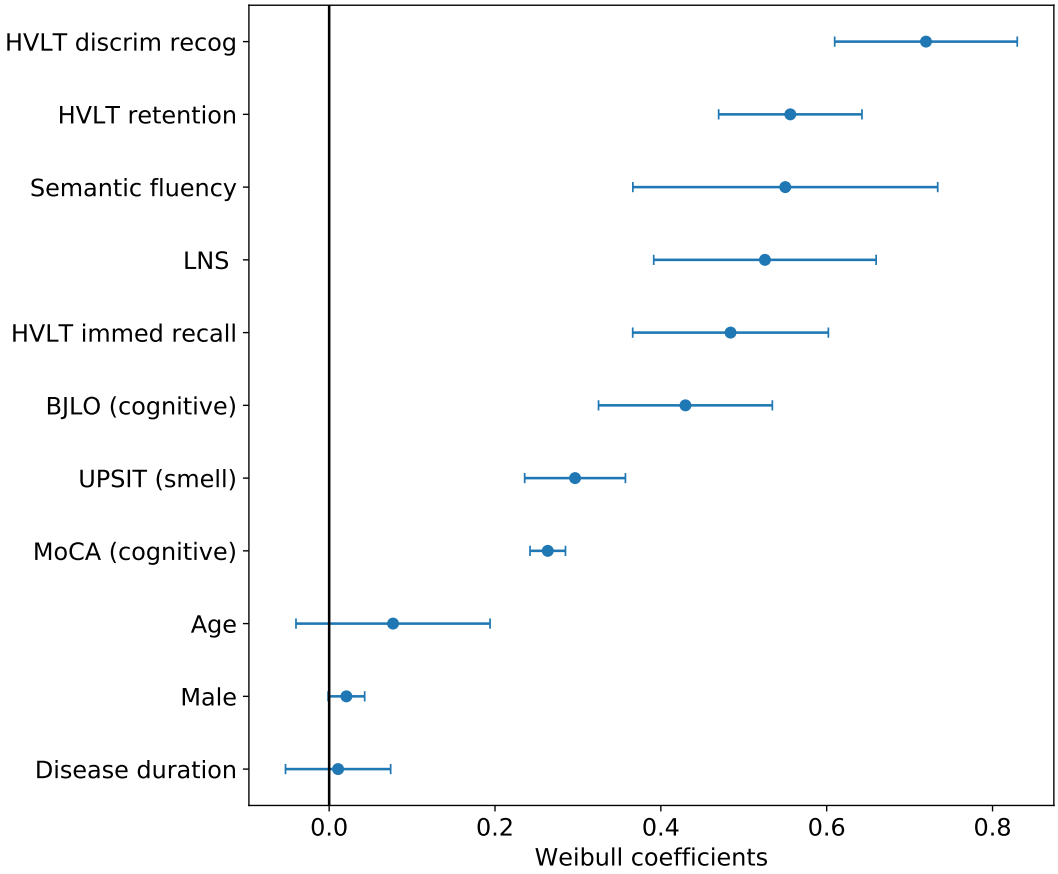

Figure 5-20: **Coefficients for cognitive outcome** from Weibull model with standard set of covariates. This model had the highest concordance index. Standard deviation across 4 folds shown.

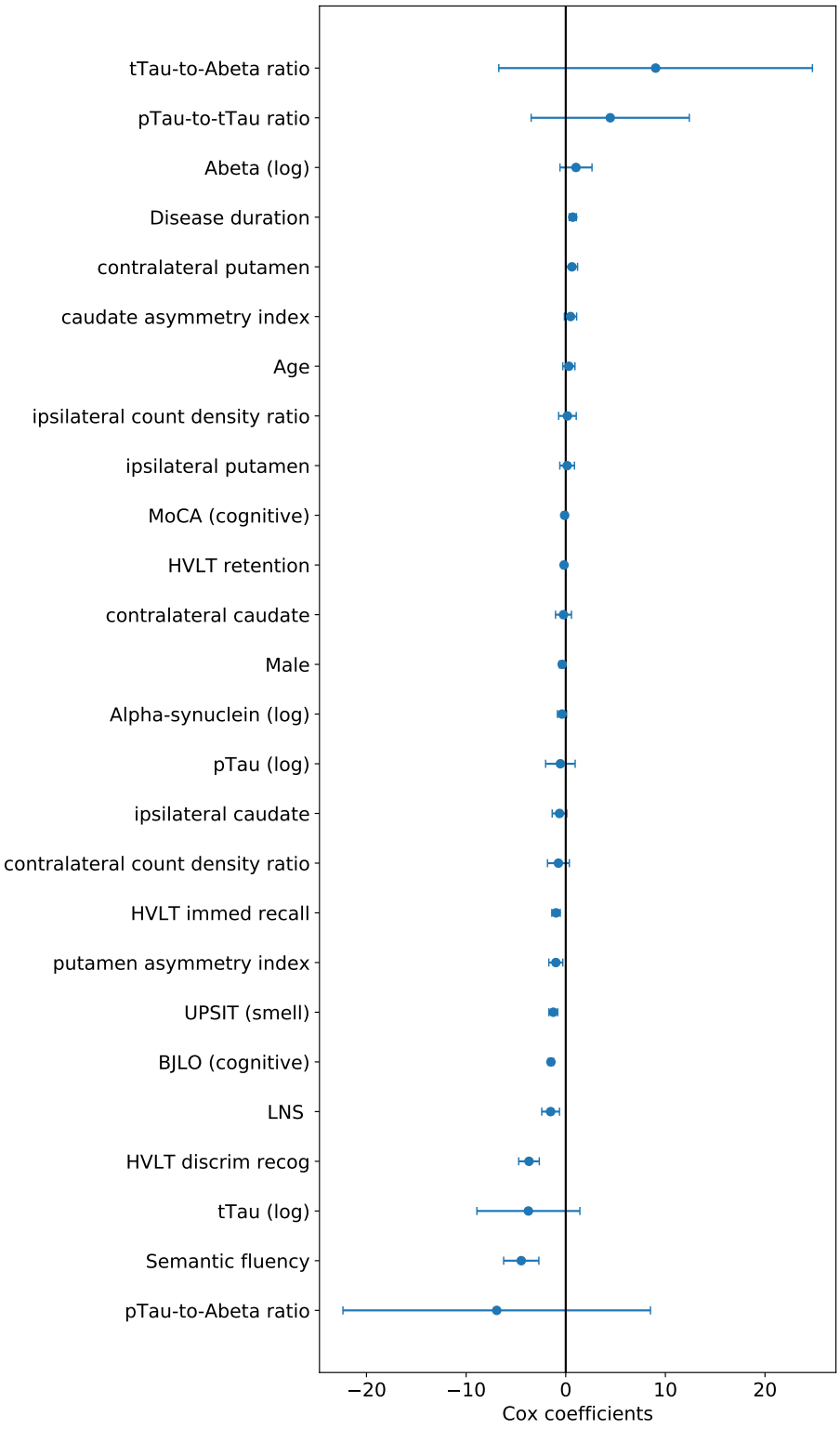

Figure 5-21: **Coefficients for cognitive outcome** from Cox model with standard + imaging + CSF set of covariates. This model had the lowest mean absolute error. Standard deviation across 4 folds shown.

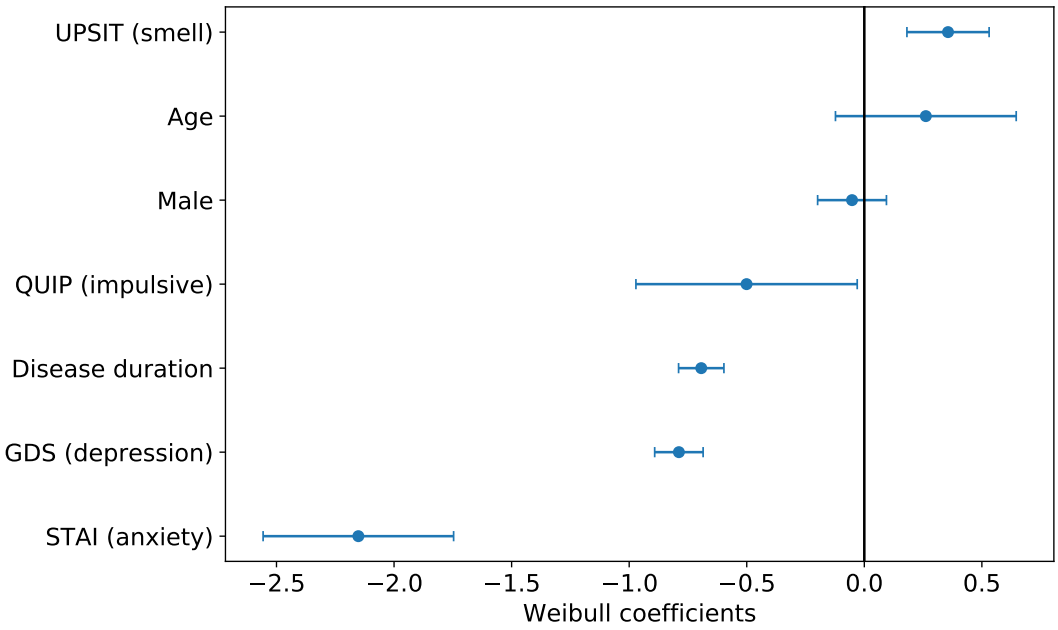

Figure 5-22: **Coefficients for psychiatric outcome** from Weibull model with standard set of covariates. This model had the lowest mean absolute error. Standard deviation across 4 folds shown.

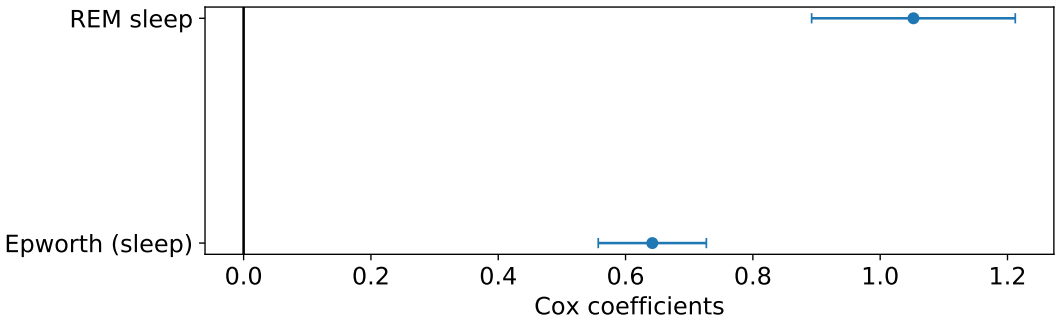

Figure 5-23: **Coefficients for sleep outcome** from Cox model with baseline set of covariates. This model had the lowest MAE. Standard deviation across 4 folds shown.

Feature	Coefficient
SCOPA drool	0.2988 (0.1331)
SCOPA constipation	0.2836 (0.1317)
SCOPA urine again w/in 2hrs	0.2297 (0.1139)
SCOPA can't completely empty bladder	0.2126 (0.0910)
MDS1 constipation	0.1974 (0.0652)
SCOPA urine weak	0.1962 (0.0906)
MDS1 light-headed	0.1961 (0.0687)
SCOPA food stuck	0.1938 (0.0658)
SCOPA male impotent	0.1874 (0.0832)
MDS1 urinary	0.1752 (0.0649)
SCOPA male unable ejaculate	0.1735 (0.0837)
SCOPA sweat excess nighttime	0.1681 (0.0581)
SCOPA difficult retain urine	0.1525 (0.0507)
MDS1 fatigue	0.1441 (0.0577)
SCOPA swallow difficult/choke	0.1284 (0.0437)
SCOPA involuntary urine	0.1045 (0.0192)
SCOPA male erection disorder med	0.0989 (0.0374)
SCOPA intolerant of cold	0.0830 (0.0137)
SCOPA urine med	0.0806 (0.0169)
SCOPA constipation med	0.0689 (0.0216)
Digestive aid	0.0609 (0.0291)
SCOPA involuntary stool	0.0566 (0.0250)
Standing diastolic BP	-0.0467 (0.0220)
contralateral caudate	-0.1322 (0.0359)
ipsilateral caudate	-0.1548 (0.0514)
ipsilateral putamen	-0.1808 (0.0790)
putamen asymmetry index	-0.2367 (0.0967)
contralateral count density ratio	-0.3429 (0.1538)

Table 5.5: **Coefficients for autonomic outcome** from Cox model with questions + treatment + imaging + CSF + expanded set of covariates. This model had the highest CI. Standard deviation across 4 folds shown in parentheses. Of the 61 features, the 28 shown here had significant coefficients.

Feature	Coefficient
contralateral count density ratio	1.0231 (0.3765)
putamen asymmetry index	0.7101 (0.2151)
tTau (log)	0.6528 (0.2300)
ipsilateral putamen	0.6257 (0.2649)
ipsilateral caudate	0.5702 (0.1297)
contralateral caudate	0.3726 (0.1590)
Supine systolic BP	0.2993 (0.0937)
Standing diastolic BP	0.2815 (0.1015)
Standing systolic BP	0.2343 (0.0631)
SCOPA urine med	-0.1821 (0.0302)
Digestive aid	-0.1850 (0.0727)
SCOPA male unable ejaculate	-0.2642 (0.0447)
SCOPA female dry vagina	-0.2680 (0.1135)
SCOPA urine weak	-0.2927 (0.1465)
SCOPA urine again w/in 2hrs	-0.3466 (0.0574)
MDS1 light-headed	-0.4301 (0.1485)
SCOPA male impotent	-0.4831 (0.1969)
SCOPA lightheaded upon standing	-0.5000 (0.2086)
SCOPA strain pass stool	-0.5054 (0.1664)
SCOPA can't completely empty bladder	-0.5232 (0.2657)
SCOPA constipation	-0.5930 (0.2062)
SCOPA drool	-0.6219 (0.1750)
SCOPA light-sensitive	-0.6234 (0.2029)
SCOPA sweat excess daytime	-0.8607 (0.3445)
SCOPA intolerant of heat	-0.8819 (0.3255)

Table 5.6: **Coefficients for autonomic outcome** from Weibull model with questions + treatment + imaging + CSF + expanded set of covariates. This model had the lowest MAE. Standard deviation across 4 folds shown in parentheses. Of the 61 features, the 25 shown here had significant coefficients.

Feature	Coefficient
injured self/partner in sleep (REM 5)	0.1481 (0.0693)
sleep when sit + read (Epworth 1)	0.1090 (0.0523)
things near bed fall in dreams (REM 6.4)	0.0924 (0.0149)
ipsilateral count density ratio	0.0768 (0.0387)
remember dreams (REM 8)	0.0529 (0.0240)
had restless leg syndrome (REM 10.d)	0.0493 (0.0233)
sleep when watch TV (Epworth 2)	0.0433 (0.0135)
had parkinsonism (REM 10.c)	0.0348 (0.0152)
ipsilateral caudate	-0.0908 (0.0393)

Table 5.7: **Coefficients for sleep outcome** from Cox model with questions + treatment + imaging + CSF + expanded set of covariates. This model had the highest CI. Standard deviation across 4 folds shown in parentheses. Of the 48 features, the 9 shown here had significant coefficients.

Although hyposmia seems unrelated to cognitive function, Gjerde et al [19] found that cognitive decline in early PD is more rapid in patients with hyposmia. The other best-performing cognitive model (Figure 5-21) also has significant coefficients on the cognitive assessments. They get dwarfed by the large coefficients on the biomarkers in the figure. It is surprising adding these biomarkers improved the mean absolute error when they have such large variance in the coefficients.

For the autonomic model in Table 5.5, several significant features are related to urinary dysfunction and constipation. In Table 5.6, light and heat sensitivity seem more important. Drooling may be an effect of poor motor control in the jaw region. The signs on the blood pressure (BP) features are consistent with how hypotension is more of a problem for PD patients than hypertension.

For the sleep outcome, both the REM sleep questionnaire and the Epworth daytime sleepiness assessment are important in the baseline Cox model (Figure 5-23). Some questions are highlighted in Table 5.7.

When predicting the psychiatric outcome, the 3 psychiatric assessments, disease duration, and hyposmia have significant coefficients, as seen in Figure 5-22. If we look at specific questions in Table 5.4, only questions related to anxiety are significant. Interestingly, only questions related to positive emotions were selected. Because some questions are very similar and may have correlated responses, that may affect the

coefficients learned. It seems odd that white race shows up as the most significant feature. 180 of the 243 white PD patients included are observed to have the psychiatric outcome, while 11 of the 12 non-white PD patients included are observed. This could just be the small non-white sample size in PPMI.

Lastly, for the hybrid outcome, the two best models both include questions in the covariate set. The model with the highest CI uses the genetic risk score, while the model with the lowest MAE does not. That is the only feature that is different, and its coefficient is not even significant. However, as seen in Tables A.2 to A.7, coefficients on the other features are vastly different between the two models.

5.5.3 Limitations

As mentioned repeatedly in the analysis above, small sample size is a strong limitation. We attempted to artificially augment the training sample size by including one sample per patient visit before the outcome and imputing missing values with the average of the preceding and successive visit. This increased sample size by approximately five-fold but significantly decreased performance (not depicted). Since the distribution of patients at later timepoints might differ from the distribution at enrollment, naïvely including later timepoints might not help with learning the distribution at enrollment. More advanced techniques for using patient trajectories might be more fruitful. It is also unclear if the outcomes are correlated and whether jointly learning the outcomes by employing a multi-task survival analysis model would help.

For the Cox proportional hazards model, an assumption is that the features satisfy the Cox proportional hazards assumption. We did not check if this assumption holds. Furthermore, the assumption might not hold linearly but instead holds when higher-order and interaction terms are included. We did not explore this avenue as the size of a reasonable covariate set is again constrained by the number of patients in the dataset.

Lastly, for a clinical decision support tool to be adopted, it must be able to achieve something that a clinician cannot notice immediately from observing a patient. A challenge with using these new outcomes in retrospective studies is that we do not

have predictions from clinicians for comparison. Thus, we cannot evaluate how much of an improvement our models can provide. Additional tasks that would be helpful include patient-specific interpretations of features driving risk prediction, adaptations for later-stage patients, and robustness checks for when these models might fail.

Chapter 6

Clinical trial sample size reduction

Clinical trials are critical for bringing potential Parkinson's disease therapies to patients. However, these trials are very expensive, in part because they require tracking large groups of patients over an extended period of time. There are three approaches to reducing the number of patients and amount of time spent tracking them: 1) identifying outcomes that are more likely to occur within a reasonable time frame, 2) identifying patients for whom these outcomes are more likely to occur when untreated, and 3) identifying patients who are more likely to benefit from the treatment. As we are not considering any particular treatment, we only demonstrate the first two approaches here. For 1), we compare the outcomes we defined in chapter 4 to state-of-the-art outcomes, such as "dead or dependent" and MDS-UPDRS total thresholds. For 2), we use two methods: a) We estimate outcome time from survival models similar to those in chapter 5 to predict if the outcome will occur during the trial period. b) We perform binary classification on early versus late outcomes using the baseline features as input. Before diving into these contributions, we start this chapter by introducing the current practice for determining clinical trial sample size.

6.1 Clinical trial sample size computation

Clinical trial proposals must include an analysis for how many patients they need to enroll to be able to reach their conclusion with sufficient statistical power. As such,

the formula most commonly used for binary outcomes is the following [56]:

$$n = \frac{(a+b)^2 (p_c * (1-p_c) + p_t * (1-p_t))}{(p_c - p_t)^2} \tag{6.1}$$

Note that n is the sample size in each of the 2 treatment groups, so twice that number of patients would have to be enrolled for a binary treatment. Before performing a hypothesis test, a **significance level** α must be set. α is the cut-off for how likely the data would have been observed had the null hypothesis been true. Thus, if the p-value is below α, the null hypothesis should be rejected. The z-statistic associated with α is a in the equation above. Typically, $\alpha = 0.05$ in a two-sided test is used, and the resulting a is 1.96. **Statistical power** in a hypothesis test measures the proportion of times the null hypothesis is correctly rejected. A clinical trial is required to have at least 80% power, and some opt for 90%. This works out to $b = 0.84$ and $b = 1.28$, respectively. p_c and p_t are the proportions of control and treated subjects, respectively, who are expected to have the outcome.

p_c can usually be ascertained from an observational study, as the control group usually receives the current practice. p_t is harder to estimate. McGhee et al [50] assume that the treatment is able to achieve a relative risk reduction of 20% or 30%. That is, $p_t = 0.8p_c$ or $p_t = 0.7p_c$, respectively. Typical time scales for clinical trials are 2 or 3 years.

6.2 Population statistics

For comparison, we also consider some outcomes that are commonly used in practice. McGhee et al [50] propose the "dead or dependent" outcome, where dependent is defined as a score below 80 on the modified Schwab and England activities of daily living scale. Because no deaths occur in the PPMI PD cohort, this outcome is driven solely by the daily living assessment. As such, we rename it the "Schwab & England" outcome. A MoCA score below 26 is considered mild cognitive impairment [55]. For the MDS-UPDRS exam, 12/13 and 32/33 are the mild/moderate cutoffs for parts II

Outcome	# pat. 2yrs	Obs 2 yrs	# pat. 3yrs	Obs 3yrs
Motor	333	35.14%	311	42.44%
MDS-UPDRS moderate	346	29.19%	322	34.47%
MDS-UPDRS severe	368	0.27%	342	0.29%
Cognitive	361	14.13%	335	17.61%
MoCA	286	25.87%	267	31.84%
Hybrid	368	25.82%	342	32.16%
Schwab & England	365	8.49%	339	11.80%
Autonomic	368	45.65%	342	49.71%
Psychiatric	368	39.95%	342	50.29%
Sleep	368	39.95%	342	47.08%

Table 6.1: **Number of patients and proportion with outcome observed** in PPMI for each outcome in 2- and 3-year clinical trial settings.

and III [49]. The same study also indicates 29/30 and 58/59 are the respective cut-offs for moderate/severe. Specifically, we compare our outcomes with their parallels in current clinical practice, specifically our motor outcome with the MDS-UPDRS outcome, our cognitive outcome with the MoCA outcome, and our hybrid outcome with the Schwab & England outcome.

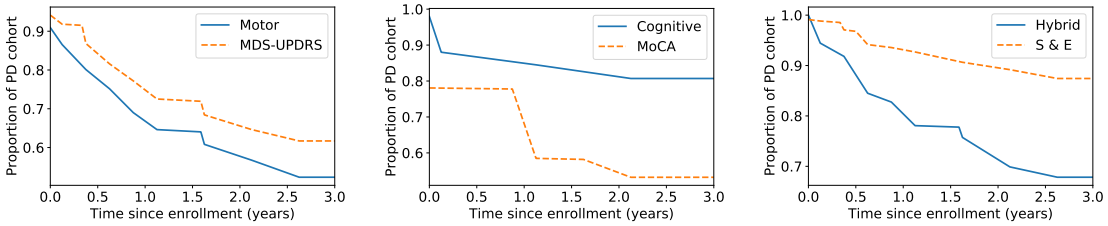

Figure 6-1: **Comparison of our novel and standard outcomes** in the 3-year filtered cohort. Standard deviation of the Kaplan-Meier curve is 0 because censored outcomes were imputed to be at the trial endpoint.

We consider only the set of 368 patients who are enrolled for at least 2 years and the set of 342 patients who are enrolled for at least 3 years. Since our outcome definitions require 2-visit persistence, sometimes the censoring time is at the second-to-last visit. To ensure that censoring times are not before the minimum enrollment time, we require that the second-to-last visit time is at least 2 or 3 years. We filter out all patients who were enrolled for less than 2 or 3 years, even if the outcome was observed when they were enrolled, to avoid any biases that might have affected the

length of time enrolled. For each outcome, we also remove patients who have the outcome at time 0 as they cannot be enrolled in a trial. In Table 6.1, we show the number of patients who are available in PPMI for each outcome and the proportion of patients who are observed with each outcome for both the 2-year and 3-year trial settings. The proportion of patients who are observed for the severe outcome defined using MDS-UPDRS is too small to be useful as clinical trial outcome for *de novo* PD patients, so we omit it from the analyses going forward.

Within the 3-year filtered cohorts, we compare the survival curves for the outcome pairs in Figure 6-1. The hybrid and motor outcomes are observed for a large proportion of the population than their counterparts. The cognitive outcome is observed for fewer patients at baseline than its counterpart. We also note that for both the cognitive and MoCA outcome extending the trial to 3 years yields very few additional observations.

6.3 Selecting patient cohorts

One way to reduce the sample size needed is to enroll patients who are more likely to have the outcome if left untreated. This would increase p_c in equation 6.1. If we assume that the relative risk reduction is the same for this enrolled cohort as for the cohort that would have originally been enrolled, then p_t would also scale accordingly. Note that this assumption may not actually hold. To increase p_c, we predict which patients are likely to have the outcome during the enrollment period. We try five types of models for this purpose: 1) Cox proportional hazards, 2) Weibull model, 3) logistic regression, 4) random forest, and 5) decision tree. As this is a binary classification problem and the last three models are binary classifiers, those three choices are intuitive. It may seem like the time-to-event prediction from survival analysis is unnecessary. However, these predictions can be treated the same way as the binary classifier outputs. If the predicted time is early, then there is a high probability the event will occur within the enrollment period. If the predicted time is much later, then there is a low probability.

We add 1 additional requirement (the third one below) to the inclusion-exclusion criteria used in the survival models:

1. No missing baseline covariates.

2. The event does not occur at time 0.

3. The second-to-last visit occurs after the trial period (2 or 3 years).

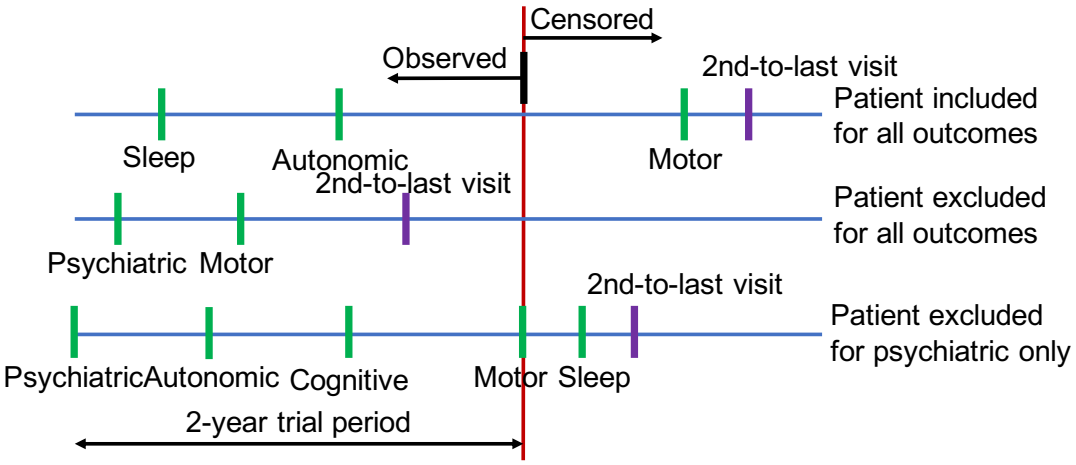

Figure 6-2: **Set-up for predicting trial outcomes** illustrated by three hypothetical patients.

To clarify this set-up, see Figure 6-2. The first patient is included for all outcomes, where sleep and autonomic are observed and the remainder are censored. The second patient is excluded for all outcomes because he/she was not enrolled for the entire trial period. The third patient is included for all outcomes except psychiatric since that occurred at enrollment. Note that any outcome that occurred at the end of the trial period, such as the motor outcome for the last patient, is considered observed.

When we split off the test set, we perform separate 80-20 splits on the observed and censored populations to ensure the test set is balanced. The 4-fold train-validation splits are similarly balanced. To avoid convergence issues when one feature has low variance in the observed or censored population, we remove binary features that have a minor frequency less than 5% in a training set in any of the 4 folds. Additionally, any feature that is removed for the cognitive outcome we designed is also removed

for the MoCA outcome and vice versa so that the cognitive models have the same features. The same relationship is held for our motor outcome and the MDS-UPDRS moderate outcome. Lastly, any feature that is removed from a model for one of the five category outcomes is also removed from the hybrid and Schwab & England outcomes.

6.4 Model performance evaluation

To evaluate these models, we measure AUROC, precision, recall, and accuracy as implemented in scikit-learn [59]. **AUROC** stands for area under the receiver operator characteristic curve, which plots the false positive rate on the x-axis and the true positive rate on the y-axis. Most models calculate a score for each sample that represents how likely the model thinks the sample is positive. By varying the threshold applied to the score, we can control how confident the model needs to be on positive samples. For AUROC, the true and false positive rates are plotted for various thresholds. AUROC is the standard metric used for binary classification.

We also focus on **precision**, which is the proportion of predicted positives that are true positives, because that is the p_c term that varies in the sample size calculation above. Recall is the proportion of positives that are predicted as positive. We also want high recall because that would reduce the screening size. There is a trade-off between precision and recall optimization. For example, an arbitrarily high precision can be achieved if the model only outputs 1 for samples it is very certain are positive. But this high precision is at the expense of low recall. Such a high precision only model would be a poor one for clinical trials as many patients who could have been enrolled from the trial are rejected. Therefore, precision and recall are typically optimized in tandem.

For survival analysis, if the predicted time is at most the trial period, the model predicts the outcome is observed. Otherwise, the prediction is that the outcome is censored. To compute the scores needed to calculate the AUROC, we subtract the observed time from the maximum predicted time and then divide by the maximum

predicted time for the score to be in the 0-1 range. Note that only the relative ordering matters for AUROC. For Cox proportional hazards, if an event is predicted to occur at infinity (i.e. never occur), the maximum predicted time is set to the maximum of the finite predicted times plus 0.5. Then the scores for those predicted to occur at infinity are 0. We note that AUROC and concordance index seem correlated as both are looking at relative orderings.

We tune the regularization settings based on validation AUROC. If multiple settings have the same AUROC, then we select the one with the best accuracy and continue with precision, recall, CI, and MAE as tiebreakers. For Cox and Weibull, we tune the same parameters as in chapter 5. For logistic regression, we use the elastic net regularizer in scikit-learn [59] and tune the L1 ratio (0, 0.5, 1) and C (1e-4, 5e-4, 1e-3, 5e-3, ..., 5e3, 1e4). For random forest and decision tree, we tune the minimum number of samples at a leaf (1, 2, 3, 4, 5, 6, 8, 10, 12, 15, 20, 25, 30, 35, 40, 45, 50, 60, 70, 80, 100). An analysis of regularization tuning is performed in Appendix A.3.

The best-performing models on AUROC and precision are specified in Tables 6.2 and 6.3. Figures 6-3 and 6-4 show the AUROC and precision of these models. We record the accuracy, recall, concordance index, and mean absolute error (last 2 only for survival models) in Tables A.8 and A.9. For example, we can see that even though both Schwab & England models 4 and 5 have precision 1, model 4 has a higher recall, making it a better model because more positive patients can be enrolled from the same size screening population.

For most outcomes, the model with the best AUROC was a logistic regression, and the model with the best precision was a Cox proportional hazards. The large variation in precision across folds was because some models did not predict any 1's and were thus assigned a precision of 0 as they cannot be used for our application. This happened for 1 fold of the 3-year best AUROC cognitive model, 1 fold of the 3-year best precision MoCA model, 1 fold of the 3-year best precision psychiatric model, 1 fold of the 2-year best AUROC cognitive model, 1 fold of the 2-year best precision Schwab & England model, and all 4 folds of the 2-year best AUROC Schwab & England model. As such, the last model cannot be used for identifying patients for

enrollment. We show some examples of logistic regression coefficients, random forest feature importances, and top decision tree branches in Appendix A.5.

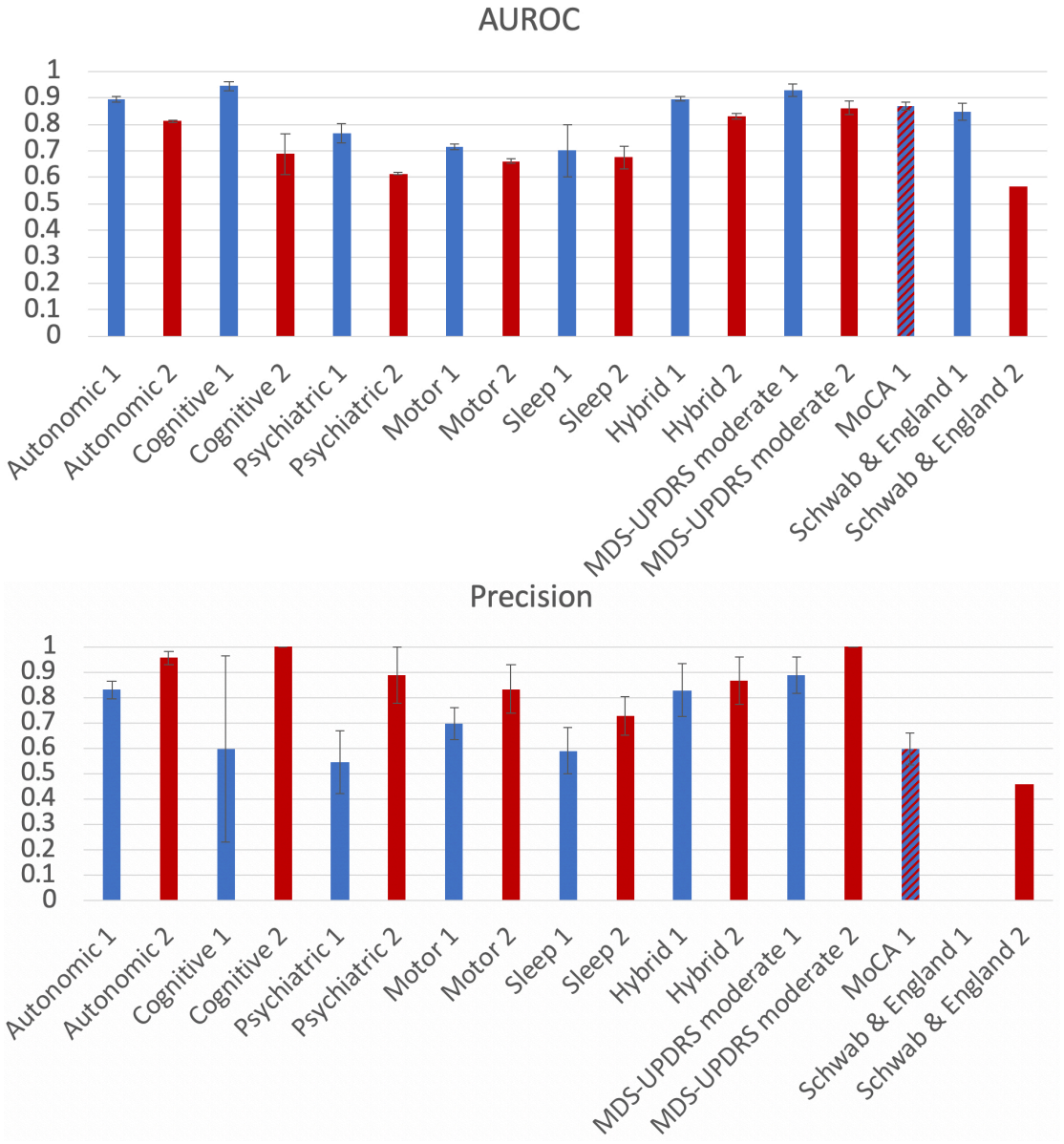

Figure 6-3: **AUROC and precision for 2-year trial setting**. Refer to Table 6.2 for the model specifications. Blue bars are models with the best AUROC. Red bars are models with the best precision. Striped bars are models that are best in both metrics. Error bars are standard deviation across 4 folds.

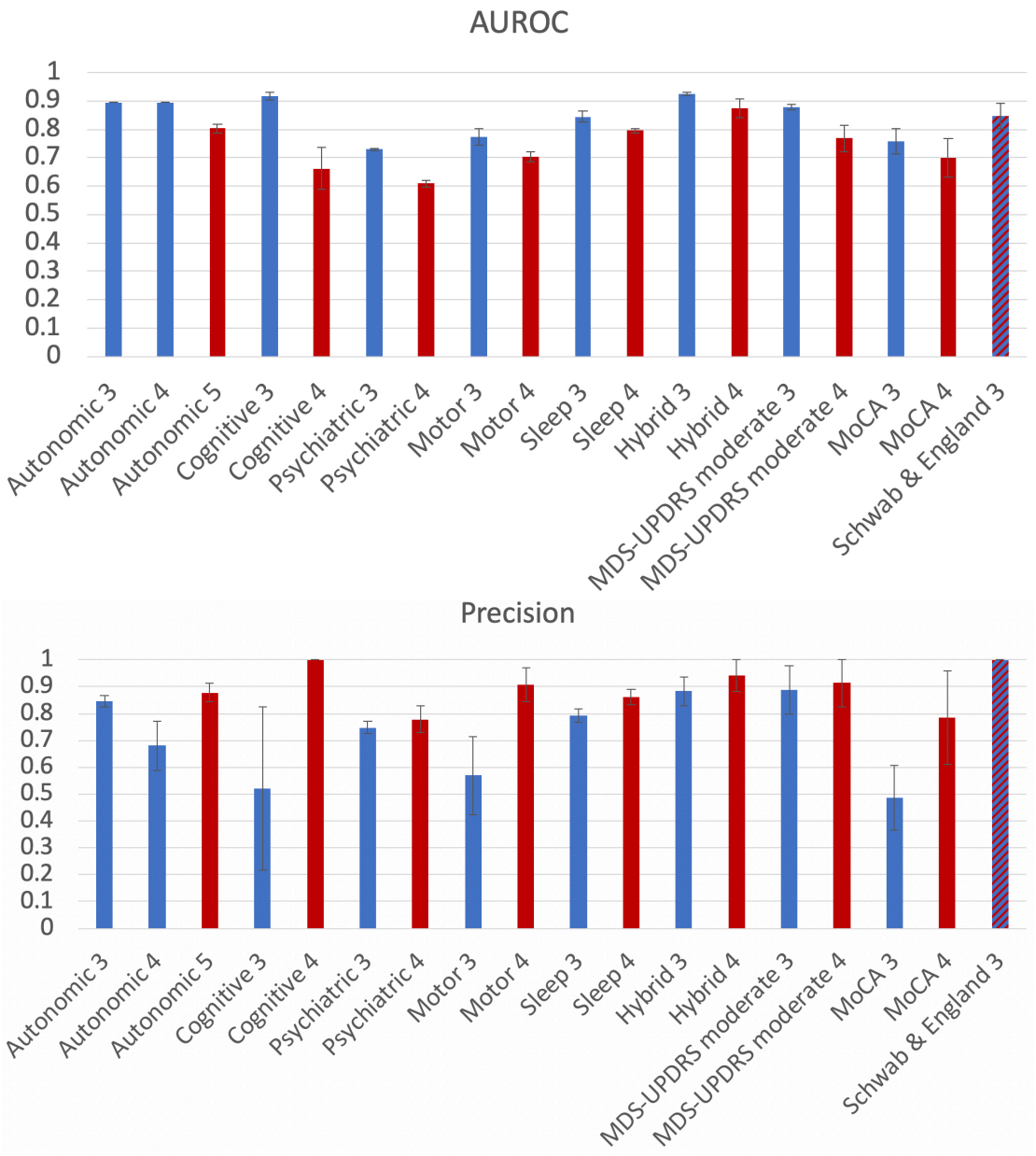

Figure 6-4: **AUROC and precision for 3-year trial setting**. Refer to Table 6.3 for the model specifications and Figure 6-3 for a description of this figure.

Outcome	Model	Covariate set	Best on
Autonomic 1	DT	Std	AUROC
Autonomic 2	CX	BL	Precision
Cognitive 1	LR	Std	AUROC
Cognitive 2	CX	Qst + Img + CSF + Exp	Precision
Psychiatric 1	WB	BL	AUROC
Psychiatric 2	CX	BL	Precision
Motor 1	LR	Qst + Img + CSF + Exp + Gen	AUROC
Motor 2	CX	MDS	Precision
Sleep 1	WB	Std	AUROC
Sleep 2	CX	BL	Precision
Hybrid 1	RF	BL	AUROC
Hybrid 2	CX	BL	Precision
MDS-UPDRS 1	LR	Qst + Img + CSF + Exp	AUROC
MDS-UPDRS 2	RF	Std + Img + CSF + Exp	Precision
MoCA 1	LR	BL	Both
Schwab & England 1	LR	Std + Trt + Img + CSF	AUROC
Schwab & England 2	CX	Std + Trt + Img + CSF	Precision

Table 6.2: **Best-performing models** for each outcome in 2-year setting. LR stands for logistic regression, RF for random forest, DT for decision tree, CX for Cox, WB for Weibull. MDS-UPDRS is the moderate MDS-UPDRS outcome. Refer to Tables 5.1 and 5.2 for the covariate set abbreviations.

Outcome	Model	Covariate set	Best on
Autonomic 3	LR	BL	AUROC (tied)
Autonomic 4	WB	BL	AUROC (tied)
Autonomic 5	CX	BL	Precision
Cognitive 3	LR	BL	AUROC
Cognitive 4	CX	Std + Img + CSF + Exp	Precision
Psychiatric 3	LR	BL	AUROC
Psychiatric 4	CX	Std + Trt	Precision
Motor 3	WB	Std + Img + CSF + Exp	AUROC
Motor 4	CX	Std + Img + CSF + Exp	Precision
Sleep 3	LR	Std	AUROC
Sleep 4	CX	Qst + Trt + Img + CSF + Exp + Gen	Precision
Hybrid 3	LR	Std + Trt + Img + CSF	AUROC
Hybrid 4	RF	Std + Trt + Img + CSF	Precision
MDS-UPDRS 3	LR	Qst + Img + CSF + Exp	AUROC
MDS-UPDRS 4	CX	BL	Precision
MoCA 3	WB	Std + Img + CSF + Exp	AUROC
MoCA 4	CX	Std + Img + CSF + Exp	Precision
Schwab & England 3	LR	Std + Trt + Img + CSF + Exp	Both

Table 6.3: **Best-performing models** for each outcome in 3-year setting. Refer to Table 6.2 for a description.

6.5 Reduction in trial sizes

We compute the sample sizes required for clinical trials with combinations of the following parameters:

- Length of trial: 2 years, 3 years
- Relative risk reduction: 20%, 30%
- Power: 80%, 90%

We assume that the standard treatments observed in PPMI would be considered control treatments for any clinical trial. This is a reasonable assumption for two reasons: 1) Since most PD medications are symptomatic rather than disease-modifying, they would often be used in conjunction with the treatment on trial. 2) Because these medications are effective at controlling symptoms, requiring patients to abstain from them would be unethical.

We also compute the trial sizes that can be obtained if we were given an **oracle**, i.e. if the true outcome was known for all patients. This allows us to see how close the sample size reductions from using our predictive models are to the maximum reduction that can be achieved.

6.5.1 Comparing our outcomes to current clinical outcomes

In section 6.2, we introduced three pairs of outcomes: our motor outcome and the moderate MDS-UPDRS outcome, our cognitive outcome and the MoCA mild cognitive impairment outcome, and our hybrid outcome and the Schwab & England outcome. By comparing the original sample sizes in Figures 6-5 and 6-6 and Figures 6-9 and 6-10, we note that using our motor and hybrid outcomes requires smaller sample sizes than their standard counterparts. The MoCA outcome is more efficient on sample size than our cognitive outcome, however. This aligns with the discussion in section 6.2 about which outcome in the pair tends to occur earlier.

6.5.2 Reductions using predictive models

Now, let us consider a fixed outcome and use predictive models to identify patients for enrollment. Models that have 100% precision are able to identify only positive samples. Thus, they are able to achieve oracle-level reduction in prediction sizes. We can see this in the 3-year best precision autonomic, 3-year best precision cognitive, 2-year and 3-year best precision MDS-UDPRS, and 3-year best precision Schwab & England models. When high precision is coupled with a reasonable recall, the screening size is also smaller than the original enrollment size. For example, for the autonomic outcome, the screening size is also cut by 77-92%, as depicted in Figure 6-11. On the other hand, the cognitive outcome has lower recall. As we can see in Figure 6-7, the screening size is only reduced by 22-45%, while the oracle is capable of reducing it by 85-89%. Unfortunately, high variability in sample sizes from models across different folds makes it harder to trust the predictions from a single model.

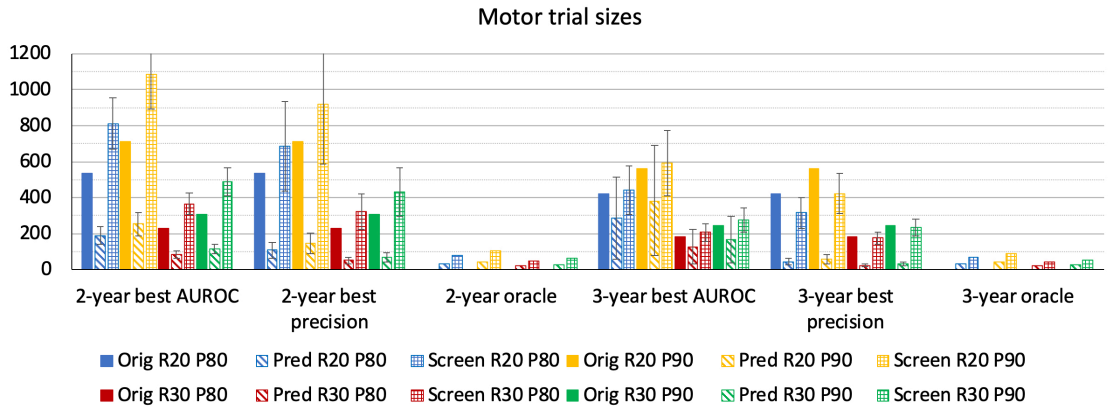

Figure 6-5: **Trial sizes for motor outcome**. Orig stands for original trial size. Pred represents the number of patients that need to be enrolled using predictions. Screen is the number of patients that need to be screened. R20 P80 means relative risk reduction of 20% and statistical power of 80%. Standard deviation across 4 folds shown for pred and screen.

Another question we can ask is if we apply these predictive models for enrollment to our outcomes and the standard outcomes, how do the enrollment and screening sizes compare then? We see that the hybrid outcome continues to maintain its advantage over the Schwab & England outcome in terms of both the enrollment and screening sizes (see Figures 6-9 and 6-10). Interestingly, the relationships flip be-

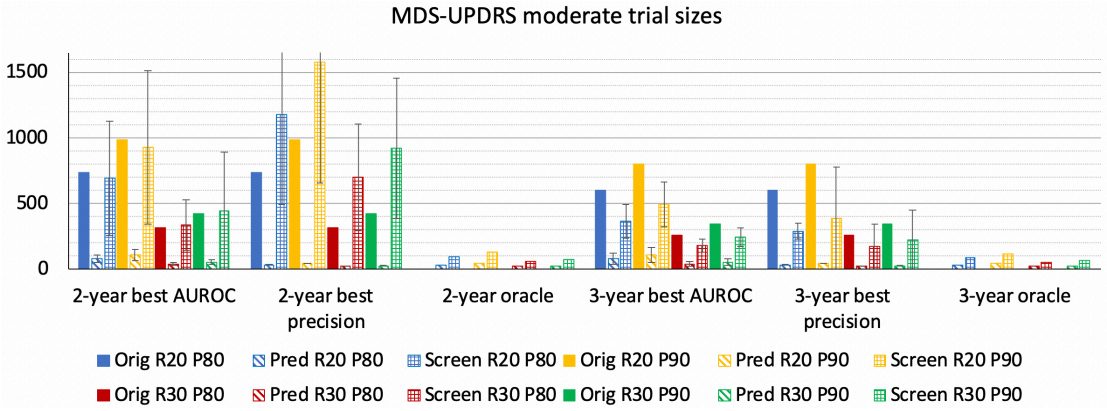

Figure 6-6: **Trial sizes for moderate MDS-UPDRS outcome**. Refer to Figure 6-5 for description.

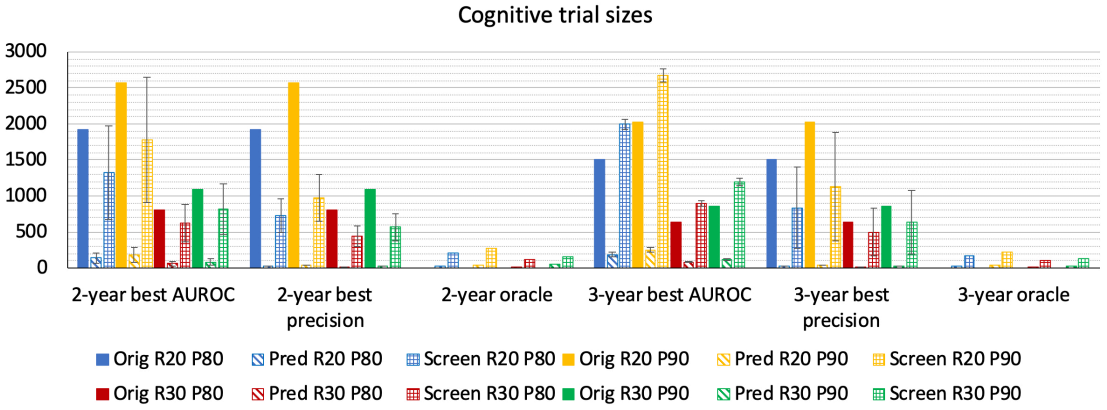

Figure 6-7: **Trial sizes for cognitive outcome**. Refer to Figure 6-5 for description.

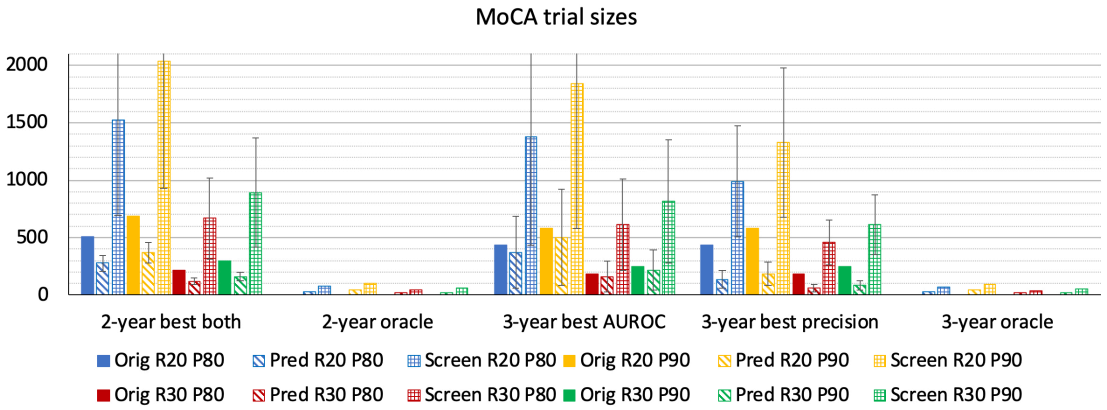

Figure 6-8: **Trial sizes for MoCA outcome**. Refer to Figure 6-5 for description.

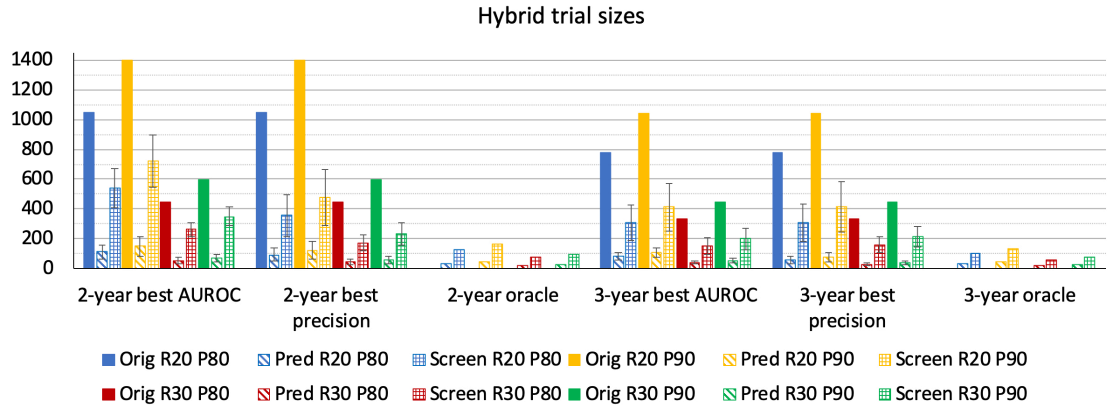

Figure 6-9: **Trial sizes for hybrid outcome**. Refer to Figure 6-5 for description.

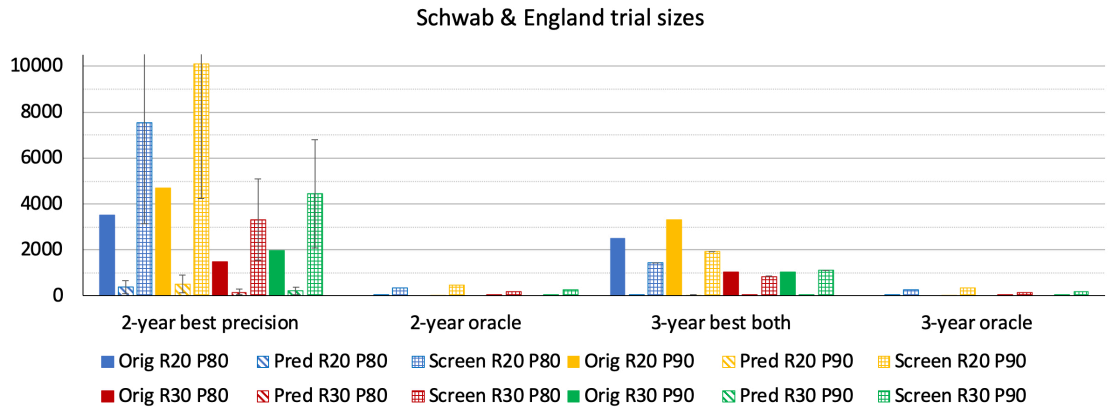

Figure 6-10: **Trial sizes for Schwab & England outcome**. Refer to Figure 6-5 for description.

tween the original and enrollment sizes for the motor and cognitive outcomes and their respective counterparts. Although the original enrollment size was smaller for our motor outcome (Figure 6-5) compared to the moderate MDS-UPDRS outcome (Figure 6-6), the moderate MDS-UPDRS outcome is easier to predict, so the enrollment size using predictions is smaller. On the other hand, when we compare Figures 6-7 and 6-8, we see our cognitive outcome now requires fewer enrolled patients than the MoCA outcome. The screening sizes are more varied depending on the setting.

In general, the best precision models seem to perform better on reducing both the enrolled and screening sizes than the best AUROC models. Because best validation AUROC was used to select the model, there is some guarantee that the recall score is reasonable, keeping the screening size in check. Overall, we see that observing patients for 3 years instead of 2 significantly reduces the number of patients that need to be enrolled for any setting. Changing the statistical power from 80% to 90% increase the number of patients required but does not have as large of an effect as changing the length of trial. An additional benefit to 3-year trials is that longer-term effects of treatments can be observed. Especially as later patient visits are spaced farther apart, it is not that expensive to collect an additional year of data per patient. However, the drawbacks are that more patients could be lost from the trial, and it will take longer from the drug to go to market. By applying our predictive models though, we can enroll fewer patients for 2 years and achieve the same statistically powerful results as the original enrollment for 3 years.

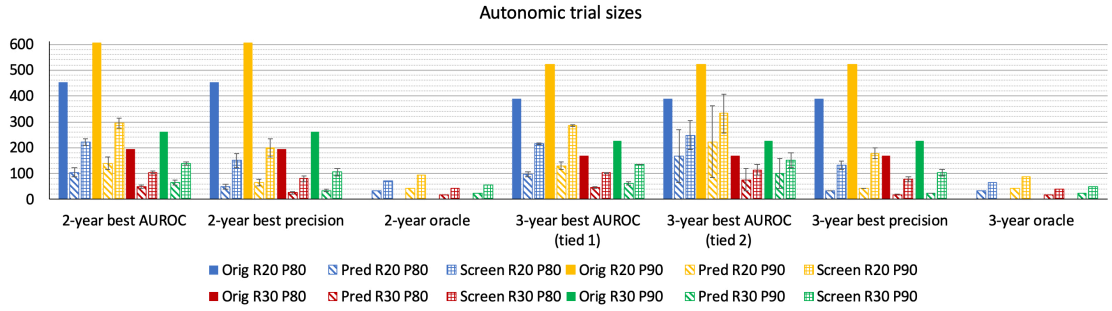

Figure 6-11: **Trial sizes for autonomic outcome**. Refer to Figure 6-5 for description.

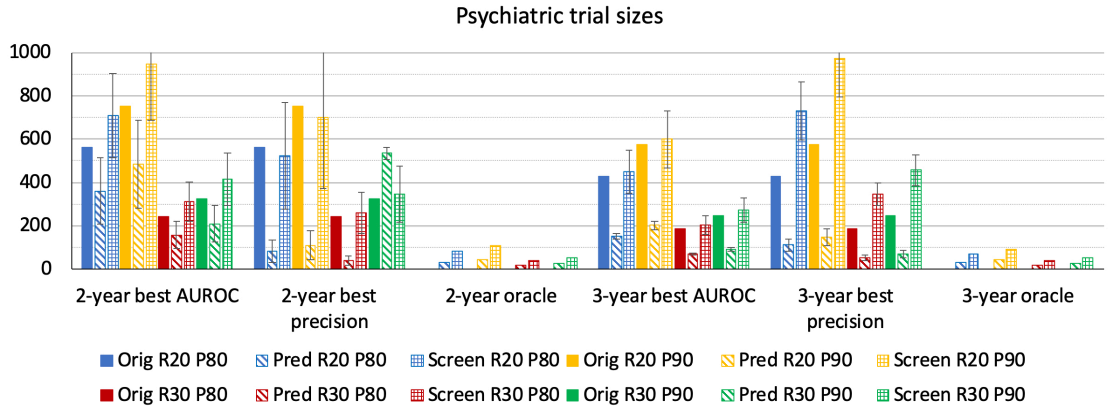

Figure 6-12: **Trial sizes for psychiatric outcome**. Refer to Figure 6-5 for description.

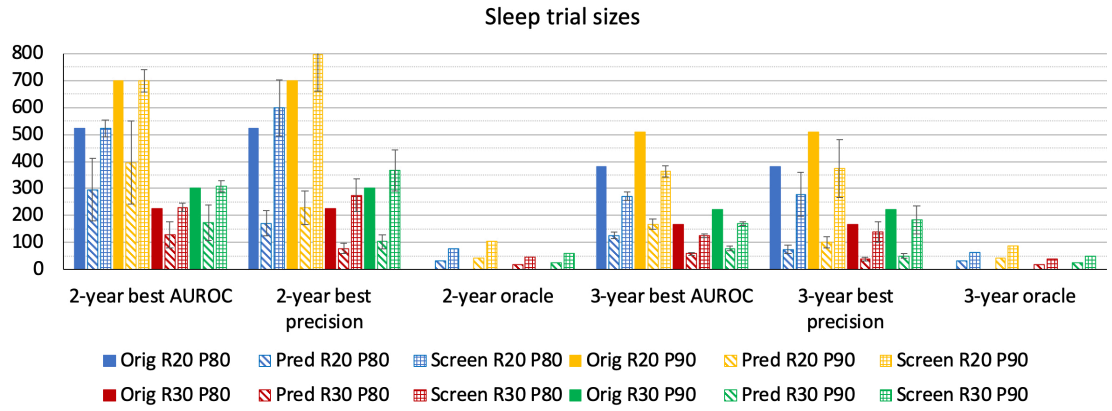

Figure 6-13: **Trial sizes for sleep outcome**. Refer to Figure 6-5 for description.

6.6 Limitations

The sample sizes are very sensitive to whether the relative risk reduction is 20% or 30%. This is the trickiest parameter to estimate as it hinges on whether the treatment being tested is effective. However, 20% or 30% seems like a reasonable choice since less effective treatments might not be worth taking. Our outcomes were defined using the *de novo* cohort. As such, they should only be used for treatments designed for early-stage PD. Otherwise, most patients will be past the outcome at enrollment time. Our results should also be verified using other PD datasets, such as Parkinson's disease biomarkers discovery [68].

We are assuming that the patients we select using the model covers the full distribution of patients who will be treated by this medication. If the medication is then given to patients who differ from those selected, there may be effects that the clinical trial did not detect. On the flip side, if the drug is only given to patients identified by the model, then other patients who could have benefited from the drug are deprived of potential treatments. One way to tackle this problem is to identify subtypes of patients. Then, if the drug is intended for specific subtypes of patients, we can apply the methods in this chapter to enroll a sufficient number of patients in each subtype who are likely to have the outcome observed during the trial period if untreated.

Chapter 7

Trajectory learning

Predicting future scores is a commonly studied clinical question. Not only can it inform patients of their future disease state, but it can also aid in other tasks, such as time-to-outcome and treatment effect predictions. Instead of predicting a score at a single timepoint in the future, we want to understand the entire trajectory of a score. As stated previously, this trajectory is driven in part by changes in disease severity and partly by noise in the assessment. In this chapter, we perform two initial forays to tackle the issue of noise, focusing specifically on MDS-UPDRS parts II and III, the standard motor assessment.

For the first foray, we model how the subtotals defined in Table 2.3 change over time. By breaking the total score into parts, we hope the noise in some questions will no longer hide progression in other symptoms. Furthermore, we can examine how treatment might relieve some symptoms but not others. Our second approach uses latent variable models to capture motor severity scores. Instead of identifying the subtotals ourselves, we allow the model to find a way to summarize the questions. Our goal is to find a representation of motor symptom severity that might be more clinically useful than the MDS-UPDRS parts II and III total. Although we have not yet reached this goal, we hope the methods we developed can help orient future work.

7.1 Learning trajectories of MDS-UPDRS subtotals

We consider some simple function forms for representing the trajectories of MDS-UPDRS subtotals. Because of heterogeneity in the population, we decided to fit separate models for each patient. This way, we can have a better picture of the distribution of trajectories among patients before deciding if we want to parameterize a population-level model.

7.1.1 Methods

Our goal is to learn a function for each patient subtotal from time since enrollment in years to the score. Because we do not have a parametric form for treatment effect, we fit separate models for each of the four treatment settings: untreated, on, off, and MAO-B only. For part II, there is only untreated and treated. As patients rarely stop treatment, any untreated timepoints after treatment initiation are excluded. Refer to Chapter 2 for an explanation of these settings.

We started with linear regressions, which have a closed form solution for the slope and intercept. As long as a patient has at least 2 visits under the treatment setting, we fit a regression to those data points. All 423 patients had at least 2 untreated exams, which follows from the PPMI enrollment criteria that treatment should not be required within a few months of enrollment. 340 patients have at least 2 on exams, 360 have at least 2 off exams, and 105 have at least 2 MAO-B only exams. We note that some patients are started on MAO-B inhibitors before other treatments, such as the second, third, and fifth patient from the top in Figure 7-2, so the MAO-B only phase typically occurs between untreated and the initiation of other treatments (first on or off exam). Not many patients are treated only with MAO-B inhibitors, and this phase tends to be very short. Ten examples of the linear trajectories we learned for the four treatment settings are shown in Figures 7-1 and 7-2.

We examine when the slopes for the different subtotals share the same sign (both non-negative or both non-positive) and how correlated they are. We find that tremor and daily activities show little agreement with the other subtotals, as seen in Figure

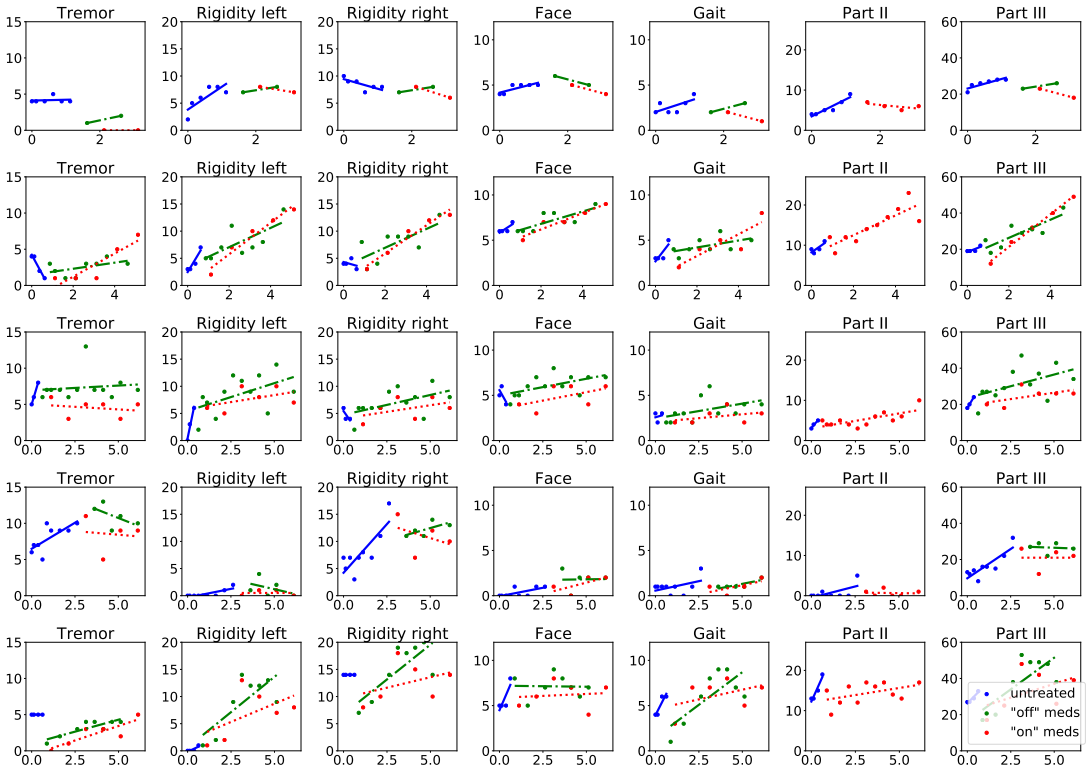

Figure 7-1: **5 linear patient trajectories learned** in PD cohort. These patients did not have MAO-B only settings.

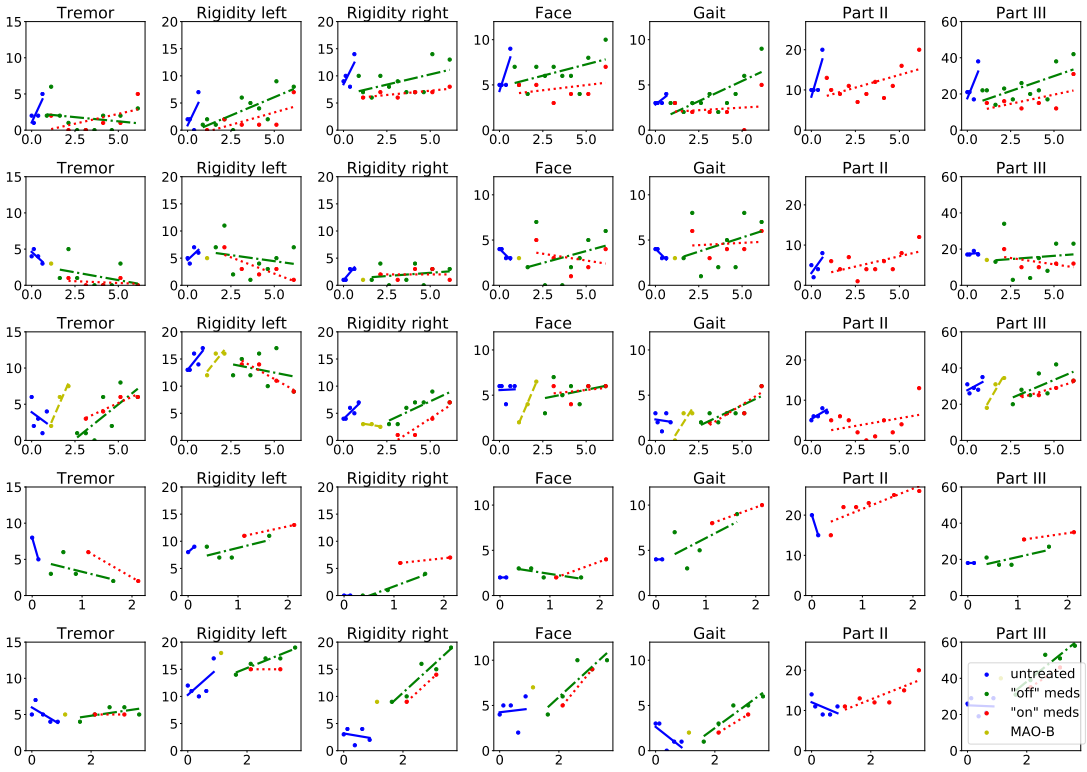

Figure 7-2: **Another 5 linear patient trajectories learned** in PD cohort.

7-3. This is particularly true for the on and off treatment settings.

It seems like some patients might experience increases or decreases in the progression rate. In particular, the linear function for on medication often ended up above the linear function for off medication. An example is the patient in the second row from the top of Figure 7-1, where the line for on medication consistently crosses above the line for off medication for all the part III subtotals. Although this could be because the on and off assessments were taken at different visits, we hypothesize that the poor fit of a linear regression might also be contributing. As such, we decided to expand the pool of function forms we considered to also include quadratic and piecewise linear models. Because a quadratic model has 3 parameters and a piecewise linear model has 5 parameters, we only use them when there are at least 3 or 5 points, respectively, under a treatment setting.

Ideally, we would fit the model to earlier timepoints and then examine the mean absolute error on the held-out later timepoints. However, because we have very few

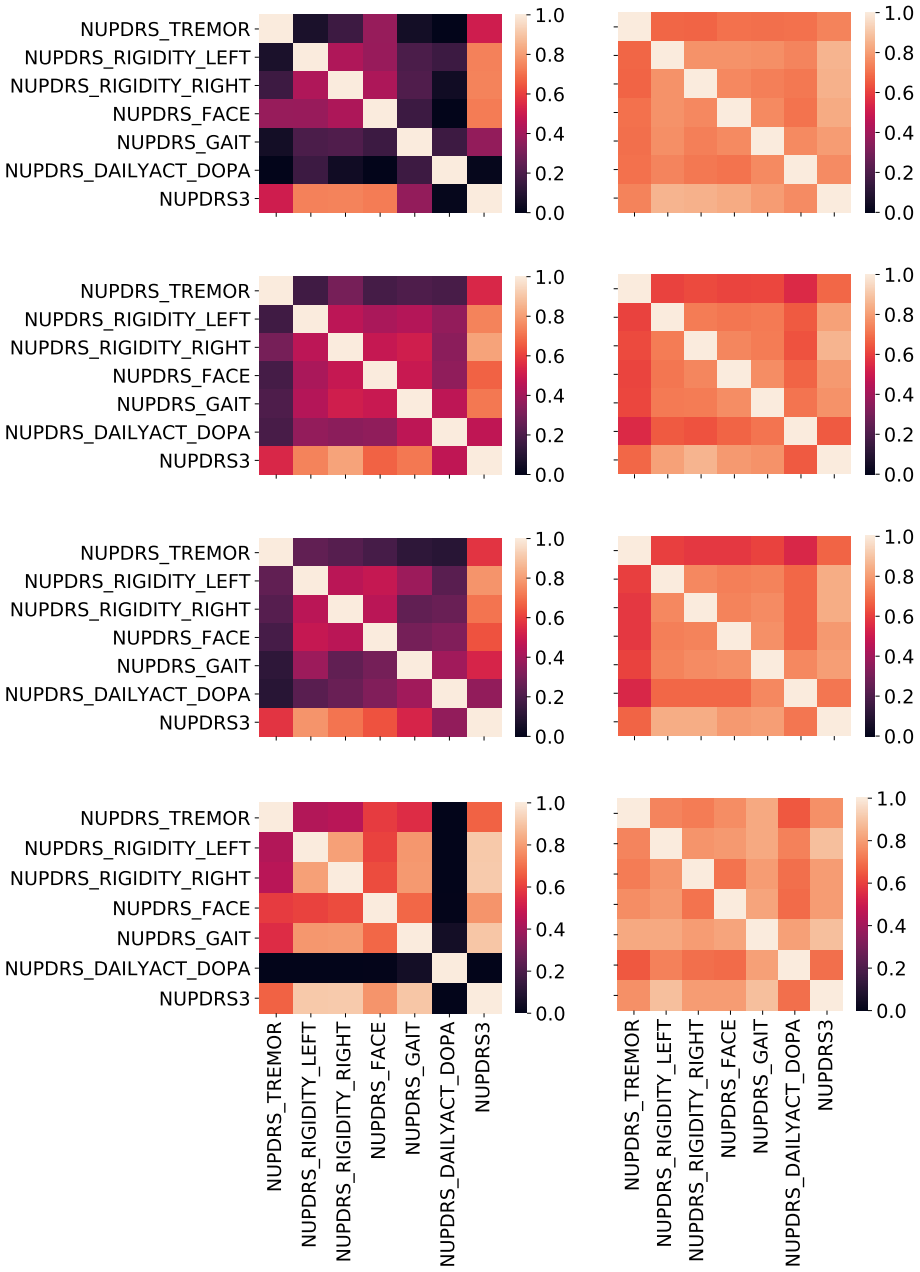

Figure 7-3: **Correlations between subtotal slopes** on left. Proportion of slopes that agree in sign (both non-positive or non-negative) on right. Top to bottom: untreated, on, off, MAO-B only. Treated subtotal for daily activities is used for the last three.

timepoints, all of them are used for fitting. We then select from the three shapes (linear, quadratic, and piecewise linear) based on which has the smallest mean absolute error on the same set of points they were fitted on. We restrict the changepoint of the piecewise linear model to be between the second and second-to-last timepoints, inclusive.

Most of the time, the on medication curves are now below the off medication curves. We can see this is the case in the ten example patients in Figures 7-4 and 7-5. These are the same ten patients as those in Figures 7-1 and 7-2, respectively. The exceptions are mostly driven by an on exam that was taken at a visit where no off exam was performed or vice versa. At least within the observation window, it seems like adding nonlinear options allows us to better represent patient trajectories.

The trajectory shapes are usually similar across the subtotals, representing uniform progression across the motor symptoms. Where the shapes differ is more interesting as that highlights heterogeneous progression. As an example, the fourth patient from the top in Figure 7-4 has right rigidity that worsens much quicker than left rigidity or any other subtotal. We can also measure the drop from the untreated trajectory to the on and off trajectories to evaluate how effective a treatment is. For example, for the second patient from the top in Figure 7-5, the treatment is particularly effective at minimizing tremor symptoms. For left rigidity, face, gait, and part II (daily activities), the medication may have initially been effective, but the scores start to rise again after some time. The continued benefits for tremor would have been masked if we only looked at the total scores.

The number of times each shape was selected for each subtotal is recorded in Table 7.1. The quadratic form was particularly common because of its flexibility. The piecewise linear model was chosen much fewer times, likely because some patients do not have at least 5 timepoints under a treatment setting.

We note that further regularization could be helpful as these functions can be particularly sensitive to a single outlier. With quadratic functions in particular, extrapolation to later timepoints could be very unwieldy. For example, for the patient in the middle row of Figure 7-4, the downward curve in the parabola for the on setting

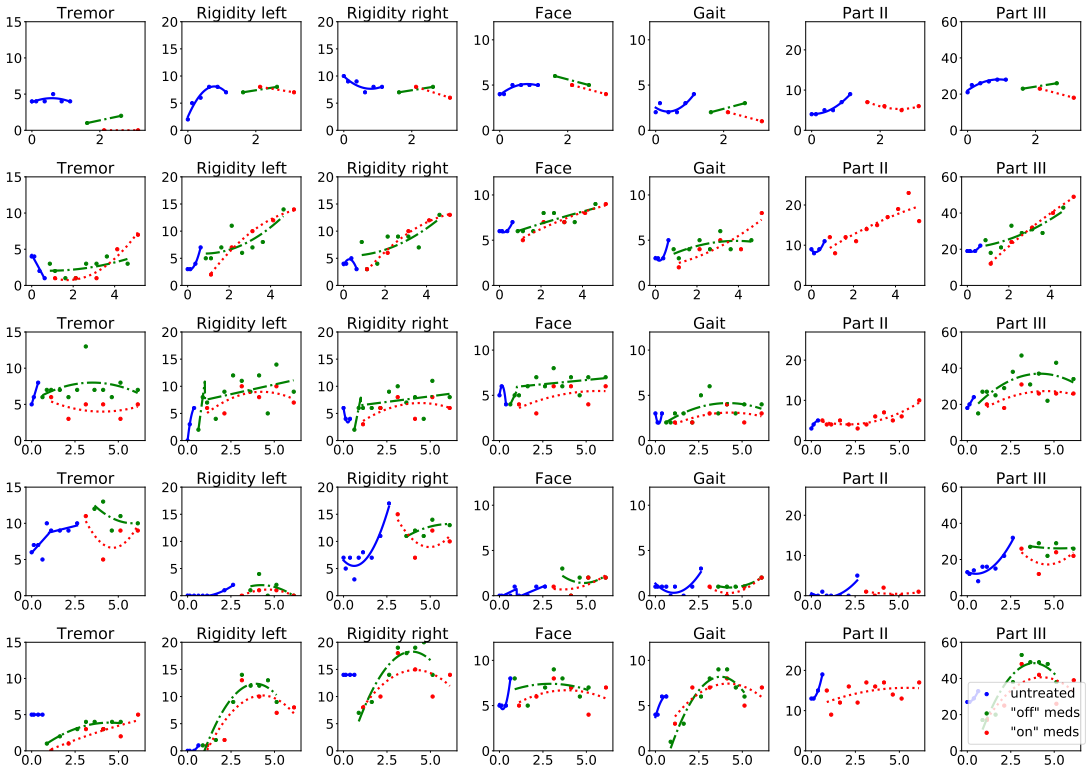

Figure 7-4: **5 nonlinear patient trajectories learned** in PD cohort. Same patients as Figure 7-1.

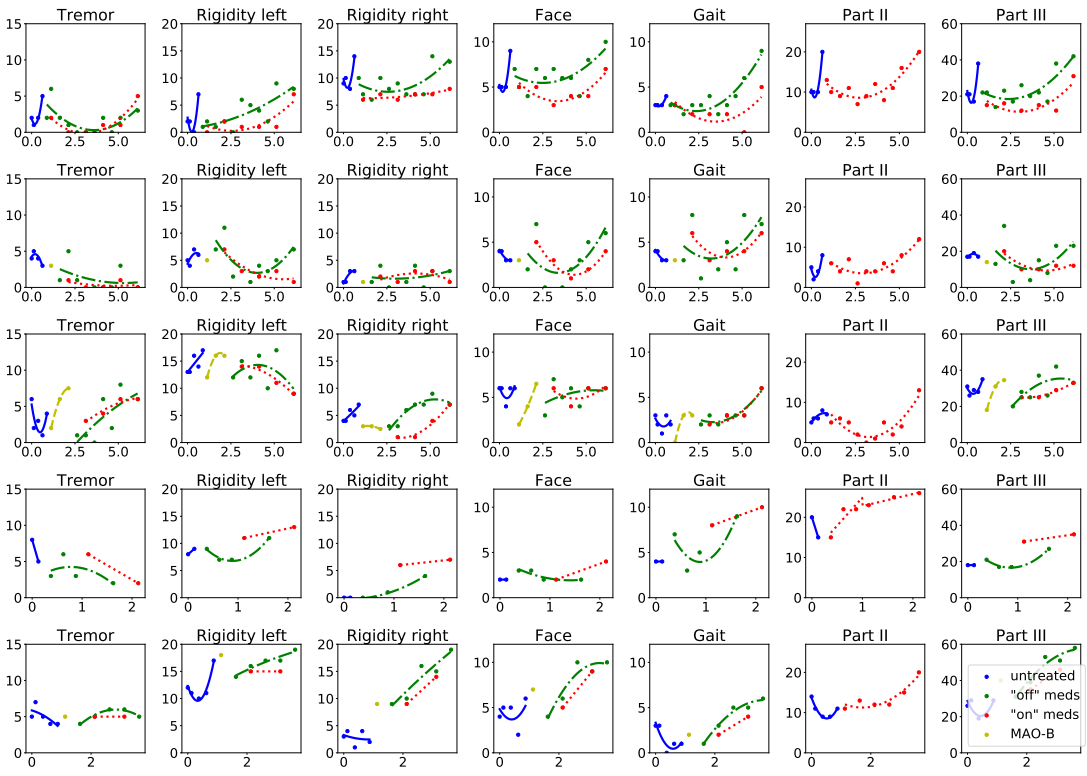

Figure 7-5: **Another 5 nonlinear patient trajectories learned** in PD cohort. Same patients as Figure 7-2.

is driven solely by the last timepoint, and it is unlikely the subtotal will continue to decrease. One reason we are particularly interested in extrapolation is to study treatment effect. One question we could ask is how the patient would have fared if he/she continued to be untreated. Estimating treatment effect by extrapolating the untreated curve and taking the difference between the projected untreated curve and the learned on or off curves would likely be problematic.

The coefficients of a quadratic form are hard to interpret. Whether a patient is near the vertex of the parabola or so far from the vertex that the model is almost linear makes a big difference. In the latter case, perhaps some regularization should be applied so that a linear model is chosen instead.

Subtotal	**Unt Q**	**Unt L**	**Unt P**	**Off Q**	**Off L**	**Off P**
Tremor	301	41	80	314	39	6
Left rigidity	300	43	77	316	35	8
Right rigidity	308	38	74	316	35	9
Face	299	40	84	308	44	8
Gait	280	63	75	297	52	7
Daily activities	359	10	15			
Part III	316	23	84	323	32	5
Subtotal	**On Q**	**On L**	**On P**	**MAO-B Q**	**MAO-B L**	**MAO-B P**
Tremor	266	66	1	61	40	2
Left rigidity	270	58	0	316	35	8
Right rigidity	278	60	0	62	40	2
Face	273	65	1	63	39	2
Gait	241	83	1	49	49	2
Daily activities	319	31	71			
Part III	286	54	0	64	40	1

Table 7.1: **Trajectory shapes for MDS-UPDRS subtotals** under the 4 treatment settings. These are counts. Q stands for quadratic, L for linear, and P for piecewise linear.

7.1.2 Future work

One of the original goals of this work was to understand if there are patterns in the trajectories that can are shared across patients in population- or subpopulation-level models. In the linear only experiment, simply examining correlation cannot give us insight into whether clusters exist. When we introduce nonlinear forms, it becomes

hard to compare trajectories using their coefficients. We can align the coefficients among the three models by setting the quadratic coefficient to 0 for the two linear models. Then, we can pretend the linear and quadratic forms are also piecewise by setting the coefficients in the two pieces to be the same and the changepoint to some arbitrary constant. However, these coefficients are still not comparable. For example, the first-order coefficient has a different meaning in a quadratic function compared to its counterpart in a linear function. If we were to use this aligned formulation, we would have more features than a clustering algorithm could handle, as there are 6 subtotals, each with 4 treatment settings and each setting has 7 parameters.

If we can detect and remove outliers before learning the trajectories, that may be helpful in conjunction with the regularization mentioned above. For example, for the first patient in Figure 7-4, the untreated tremor score in the top left corner is almost constant at 4, and the curve in the parabola was driven by solely the single score of 5. We again see how an increment of 1 might affect the function learned in the face score for the 4th patient from the top of that figure. A function that is constant at 0 and then constant at 1 may be more accurate than the two increasing pieces obtained because of the first score of 1.

Another nonlinear form that is commonly used in pharmacodynamic models is the logistic function since it is better at representing bounded growth. We note, however, that the *de novo* patients have not reached the maximum possible score. To unify the four treatment settings into a single model, we can turn to common function forms for the treatment term in pharmacodynamic models, such as those in Venuto et al [85]. Further work on understanding how to relate on and off settings is needed. It may be helpful to process the treatment dosage information and consider treatment complications such as those recorded in MDS-UPDRS part IV for this part.

7.2 Latent variable models

In the previous section, we represent the patients using pre-defined MDS-UPDRS subtotals and learn individual-specific models. In this section, we leverage the popu-

lation data to learn a new patient representation. We use latent variable modeling, which was introduced in section 3.5. The work in this section was done for a class project in collaboration with Suchan Vivatsethachai and Liyang (Sophie) Sun.

Figure 7-6 outlines the workflow for this section. First, we train various autoencoders to learn latent factors of input features. Then, we estimate progression rate and extrapolate to future time points. Table 7.2 outlines the 7 types of autoencoders we use. The baseline linear factor analysis is implemented using the factor_analyzer package [32], where the number of latent factors is selected to be the number of eigenvalues greater than 2. The remaining methods are detailed below.

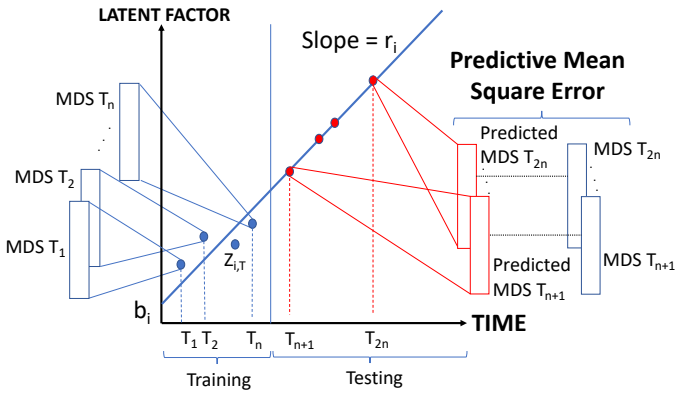

Figure 7-6: **Workflow overview for latent variable models**

Method/Property	**Encoder**	**Decoder**	**Train w/ LON**
Linear FA	Linear	Linear	No
VAE	Nonlinear	Nonlinear	No
mVAE CS	Nonlinear	Monotonic polynomial	No
mVAE LON	Nonlinear	Monotonic polynomial	Yes
Linear ordinal CS	Linear	Ordinal	No
Nonlinear ordinal CS	Nonlinear	Ordinal	No
Nonlinear ordinal LON	Nonlinear	Ordinal	Yes

Table 7.2: **Overview of methods**. FA: factor analysis. mVAE: VAE with monotonic constraint. CS: cross-sectional. LON: longitudinal. Train w/ LON refers to training with a longitudinal loss function.

We only consider the untreated timepoints to avoid having confounding due to treatment. We elect not to model the other treatment settings because less data is available for them and the different treatment initiation times and dosages introduces

too much variability. We have an average of 5.3 untreated exams per patient, with a standard deviation of 2.4.

7.2.1 Variational autoencoders with a monotonicity constraint

As described in section 3.5, Pierson et al [60] developed a model to infer multi-dimensional rates of aging from cross-sectional data. To recapitulate, they have a low-dimensional latent state that evolves linearly and map this latent state to the features using a nonlinear function. Their model is similar to a variational autoencoder (VAE) with an additional constraint that some of the features change monotonically with some amount of noise. We believe such a model might also work well for modeling PD patients because PD symptoms do not improve over time without treatment.

The dataset used in Pierson et al [60] has only one timepoint per patient. For a small subset of patients, there are two timepoints. They add a term to their loss function for minimizing the prediction loss on the second timepoint. To utilize the longitudinal data in PPMI, we consider two approaches: 1) We treat data points from the same patient as independent and apply the original model. 2) We modify their objective function to incorporate multiple longitudinal timepoints while learning. We refer to our modification as the longitudinal monotonic VAE model (mVAE LON) and their original as the cross-sectional monotonic VAE model (mVAE CS). We also use a regular VAE as a nonlinear baseline. We tune the encoders and VAE decoder as neural networks with up to 3 hidden layers and 2, 10, or 20 hidden units in each hidden layer. The monotonic decoder function is a polynomial. As such, we vary the polynomial powers. For example, one setting we tried was having terms with powers 0.2, 0.5, 1, 2, and 3. Because the mVAE Is not robust, we try 10 random initializations and report the best results here.

VAE-based models require large amounts of data, but we have less than 2000 points in our training set. Thus, we synthesize intermediate timepoints via linear interpolation of adjacent timepoints. This also helps for the longitudinal model since it provides any timepoints used in the loss function that might be missing. Interpolation also allows us to obtain multiple samples per patient by taking a sliding window along

the patient's visits. Assume that for patient i, we observe features of D questions $(x_{d,i,t} \mid d = 1, \ldots, D)$ for time $t = t_1, \ldots, t_T$. To sample a data point for time t' such that $t_l < t' < t_{l+1}$, we approximate the score $x_{d,i,t'}$ by linear interpolation from the adjacent observed score x_{d,i,t_l} and $x_{d,i,t_{l+1}}$. Specifically, we set

$$x_{d,i,t'} = x_{d,i,t_l} + (t' - t_l)\frac{x_{d,i,t_{l+1}} - x_{d,i,t_l}}{t_{l+1} - t_l} \tag{7.1}$$

for all $d = 1, \ldots, D$. An example of interpolated scores for sampled data points is illustrated in Figure 7-7.

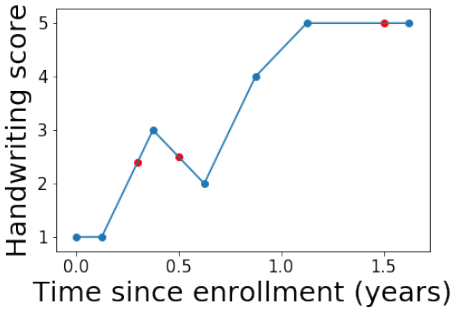

Figure 7-7: **Example of data augmentation** for a patient's handwriting score. The blue dots are the scores of the real observed data point. The red dots are the interpolated scores of the sampled data points.

7.2.2 Ordinal regression

In a typical auto-encoder, both the latent factors and the features are continuous. In our case, the question responses that serve as the features are discrete. Treating them as continuous features may result in inaccurate predictions. The alternative–treating discrete features as categories in a multi-label classification problem–does not recognize some categories are closer than others. Ordinal regression models map continuous inputs to discrete outputs by learning thresholds to separate the outcomes, thereby capturing this ordering. The loss function is similar to a classification problem, where the penalty is one-sided, i.e. it is applied only if the prediction is on the wrong side of the threshold. Rennie et al [67] demonstrate that penalizing all erroneous thresholds is better than just penalizing the nearest incorrect threshold. Thus, we use their

all-threshold one-sided mean squared error in our objective.

Our evaluation on two types of metrics (prediction error and ranking of latent factor by time) closely parallels the paradigm in survival analysis, where the two aims are predicting time to observed events and ordering censored events. Thus, we borrow a loss function that balances these two objectives from Liu et al [40]. Putting these elements together and modifying the model to set thresholds for multiple ordinal questions using the same latent factor, we obtain the following objective function:

$$l(x,y) = \sum_{d,i,t} \sum_{k=1}^{K} \left((z_{i,t} - \theta_{d,i,t})^2 \left(1\left\{ y_{d,i,t} \geq k, x_{d,i,t} < k \right\} + 1\left\{ y_{d,i,t} < k, x_{d,i,t} \geq k \right\} \right) \right) - \frac{\alpha}{\sum_i |\mathcal{E}^{(i)}|} \left(\sum_i \sum_{(p,q)\in\mathcal{E}^{(i)}} \log \sigma \left(z_{i,p} - z_{i,q} \right) \right), \tag{7.2}$$

where d index the questions, i the patients, and t the visit number. $x_{d,i,t}$ denotes the input observed features, while $y_{d,i,t}$ denotes the output features. k is the discrete response, which ranges from 0 to K and the thresholds are θ_{dk}. σ is the sigmoid function, α is a non-negative constant, and $\mathcal{E}^{(i)}$ is the set of all pairs of ordered timepoints, i.e. existence of an edge in $\mathcal{E}^{(i)}$ implies time index $p > q$ for patient i. We vary α to tune the weight of the ranking loss. Setting $\alpha = 0$ gives us the cross-sectional model.

7.2.3 Evaluation metrics

First, we examine the latent factors our models discover by studying how they correlate with the observed features. Second, we consider four quantitative metrics that measure how informative the latent factors are to the PD progression. We measure how well the latent factors obey time ordering using the concordance index (CI) and consecutive visit ranking (only adjacent visit pairs are considered). An ideal latent representation would be monotonic with respect to time since it would capture disease severity without noise. For models with multiple latent factors, we calculate CI

for each latent factor, but only report the highest here, assuming that other latent factors may be capturing non-temporal information. For the longitudinal model, we have multiple sliding windows per patient, so we compare the latent factor at the first timepoint of each window.

We also evaluate how well the autoencoders can reconstruct input features. Specifically, we calculate the mean squared error (MSE) between the reconstructed and observed features. For prediction, we first fit a patient-specific linear regression of latent factor on time since enrollment. The first half of timepoints are used for fitting; the second half for prediction. Feeding the extrapolated latent factors into the decoder outputs predictions for the question responses. MSE can then be calculated between the predictions and observations. Note that MSEs for ordinal models are necessarily larger since ordinal models cannot output non-integer values.

7.2.4 Results and discussion

As seen in Figure 7-8, the latent factors show a generally increasing trend but with noisy fluctuations. Figure 7-9 shows that when there are one or two latent factors, almost all the features are positively correlated with them, as all the MDS-UPDRS II and III questions are generally correlated. With multiple latent factors or hidden units, the VAE is best at disentangling as expected, but the features related to each latent factor do not form coherent subgroups.

Table 7.3 shows that the VAE performs better than the linear baseline on MSEs, indicating that nonlinearity is indeed necessary. The cross-sectional ordinal models perform best on the ranking metrics since a thresholding function is monotonic. The cross-sectional version of the monotonic VAE performs the best all-around, as it has the best prediction MSE and is close to best on the other metrics. This is likely because it captures the benefits of nonlinearities as in the VAE and monotonicity from its polynomial decoder. The decreased performance of both longitudinal models requires further analysis.

Because the latent factor is correlated with time since enrollment, we can interpret question responses with lower thresholds as symptoms that tend to occur earlier for

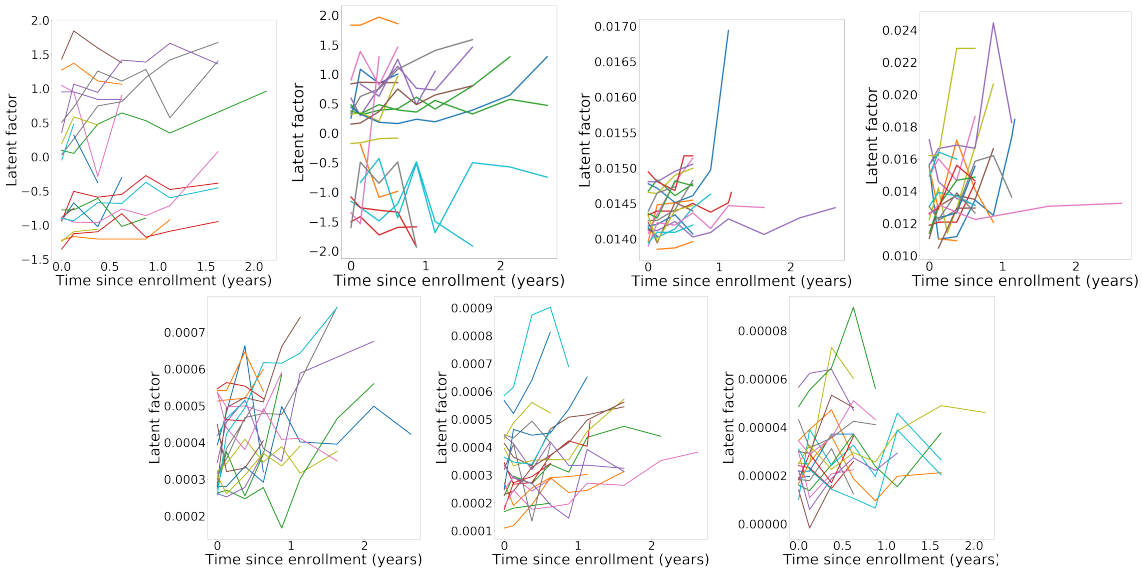

Figure 7-8: **Latent factors across time** for 20 randomly sampled patients in test set. Left to right: Top row: First latent factor in linear FA. VAE. mVAE CS. mVAE LON Bottom row: Lin ord CS. Nonlin ord CS. Nonlin ord LON.

Method/Metric	CI	Consec. rank	MSE recons.	MSE pred.
Linear FA	0.682 (0.010)	0.577 (0.006)	0.862 (0.010)	1.533 (0.030)
VAE	0.570 (0.026)	0.528 (0.016)	**0.392** (0.016)	1.344 (0.538)
mVAE CS	0.727 (0.036)	0.616 (0.020)	0.559 (0.016)	**0.631** (0.018)
mVAE LON	0.657 (0.031)	0.561 (0.013)	0.666 (0.084)	0.846 (0.107)
Lin ord CS	0.730 (0.026)	**0.619** (0.028)	1.504 (0.015)	1.856 (0.089)
Nonlin ord CS	**0.739** (0.017)	**0.619** (0.025)	1.302 (0.013)	1.582 (0.051)
Nonlin ord LON	0.651 (0.011)	0.554 (0.011)	1.897 (0.038)	2.163 (0.078)

Table 7.3: **Performance of methods** on test set in 5-fold cross-validation. Mean from 5 folds is shown, followed by standard deviation in parentheses. Best model on each metric is bolded in its column. Consec. rank: consecutive visit ranking. MSE recons.: MSE on reconstructing input data. MSE pred.: MSE on predictions.

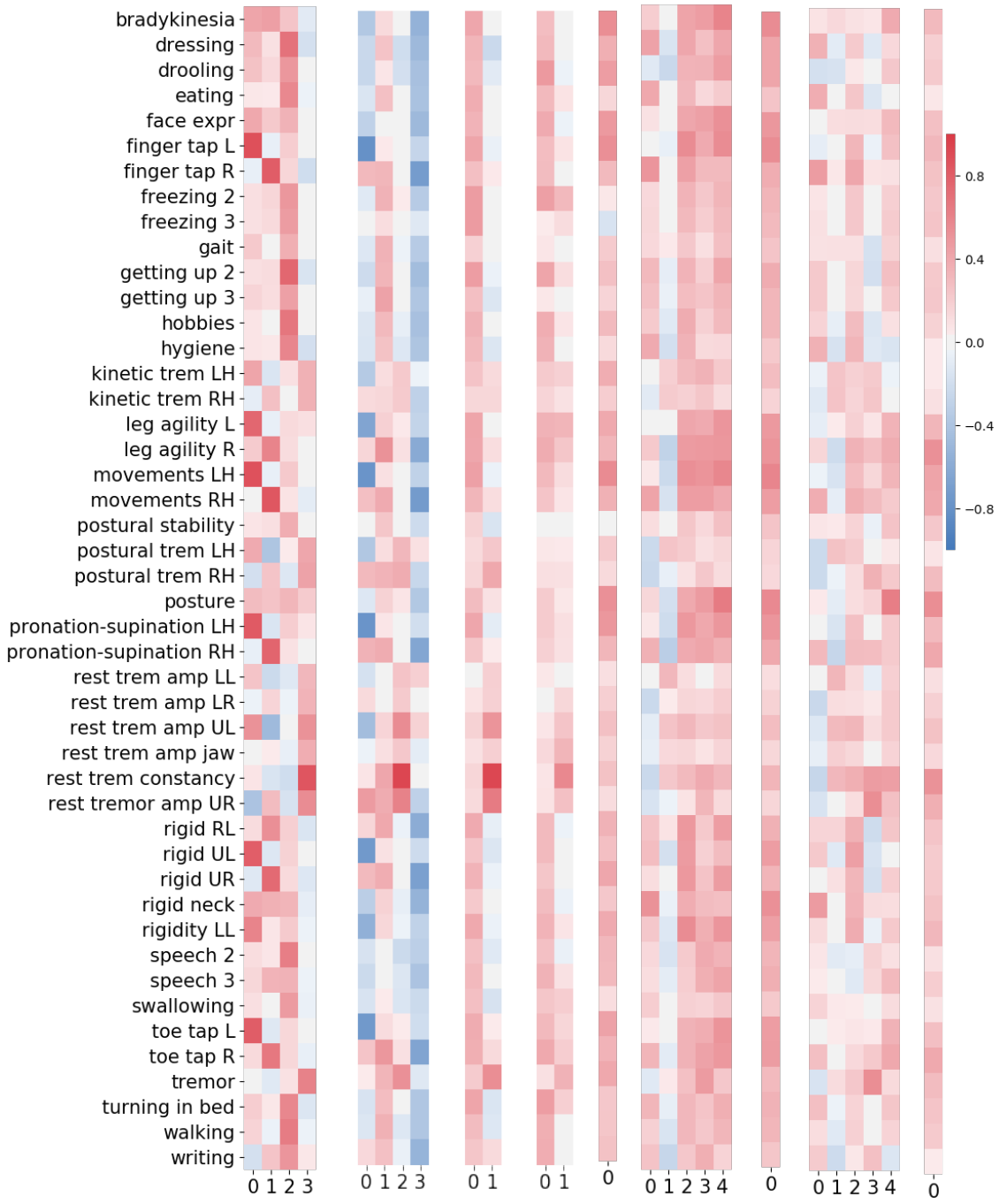

Figure 7-9: **Correlation between latent factors and observed features** in each model. Left to right: Linear FA. VAE. mVAE CS. mVAE LON. Lin ord CS. Nonlin ord CS hidden. Nonlin ord CS latent. Nonlin ord LON hidden. Nonlin ord LON latent.

patients. For example, as seen in Figure 7-10, larger rest tremor amplitude in the upper left body and rigidity in the lower right body might appear earlier than difficulty getting up from a deep chair, impairment of hobbies, or postural tremor in the right hand.

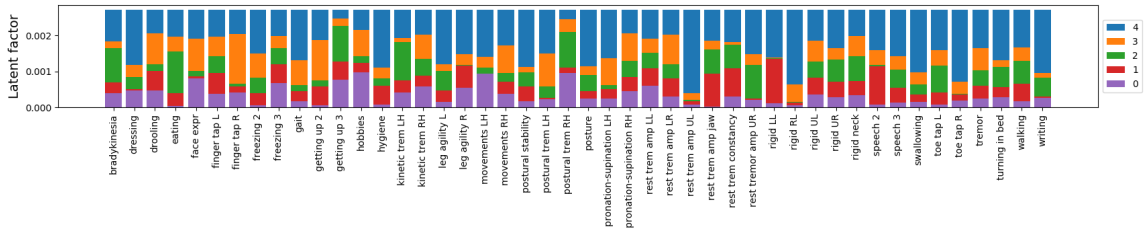

Figure 7-10: **Ordinal thresholds** from nonlinear cross-sectional model.

Unfortunately, none of our models had higher concordance indices than the MDS-UPDRS parts II and III total itself (0.7478). Because we would expect motor severity to be monotonic, we still have some work to do if we want to have a less noisy representation of motor severity. It would also be interesting to evaluate how well we can reconstruct and predict the question response from the total score to compare the MAE metrics.

A limitation of the deterministic single latent factor ordinal model is that it cannot model noise in question responses and satisfy the constraints imposed by multiple questions. Another limitation that applies to all the methods in this section is that the data distribution, specifically the relationship between latent factors and observed features, shifts across time. Because all patients in the *de novo* cohort are within a few years of diagnosis, this is not a problem in the dataset we test on. However, this may be an issue if we apply our model to later-stage PD patients.

We can address the limitation above for ordinal models by using multiple latent factors, where each latent factor is tied to a subgroup of related questions. We can also implement a probabilistic model. More broadly, because most PD patients eventually start treatment, we should modify all our models to account for treatment. This would vastly increase the clinical utility of our model. We can also expand the set of input features to include other assessments. Finally, validating our findings on another PD dataset and applying our methods to other diseases or longitudinal

survey datasets can broaden our impact.

If the trajectories of multiple features are related by some underlying disease etiology, then we may be able to leverage the benefits of learning these trajectories together by using machine learning methods. Formalizing our work in section 7.1 as a prediction problem can serve as a baseline for these more complex models. The work in section 7.2 presents one method for learning multiple trajectories that can be applied to more diverse settings, such as assessments across different categories of symptoms. As this chapter is most akin to the disease progression modeling work discussed in sections 3.1 and 3.2, we hope this can steer future models of PD progression towards developing better patient representations.

Chapter 8

Subtyping progression patterns

For a patient, knowing how other patients who are similar to them progress can provide a reference for themselves. This can help guide decisions on whether they should start treatment or make provisions for accommodations later in life. For subtypes to be meaningful this way, they must capture future prognosis. We believe that machine learning methods can help in this endeavor because they are more capable of handling longitudinal data and a larger number of features. To this end, we consider three approaches shown in Figure 8-1: 1) using the latent factor models from section 7.2, 2) performing non-negative matrix factorization, and 3) clustering the outcomes from chapter 4. The unifying theme behind these approaches is that each captures a low-dimensional representation of heterogeneity in the whole trajectory.

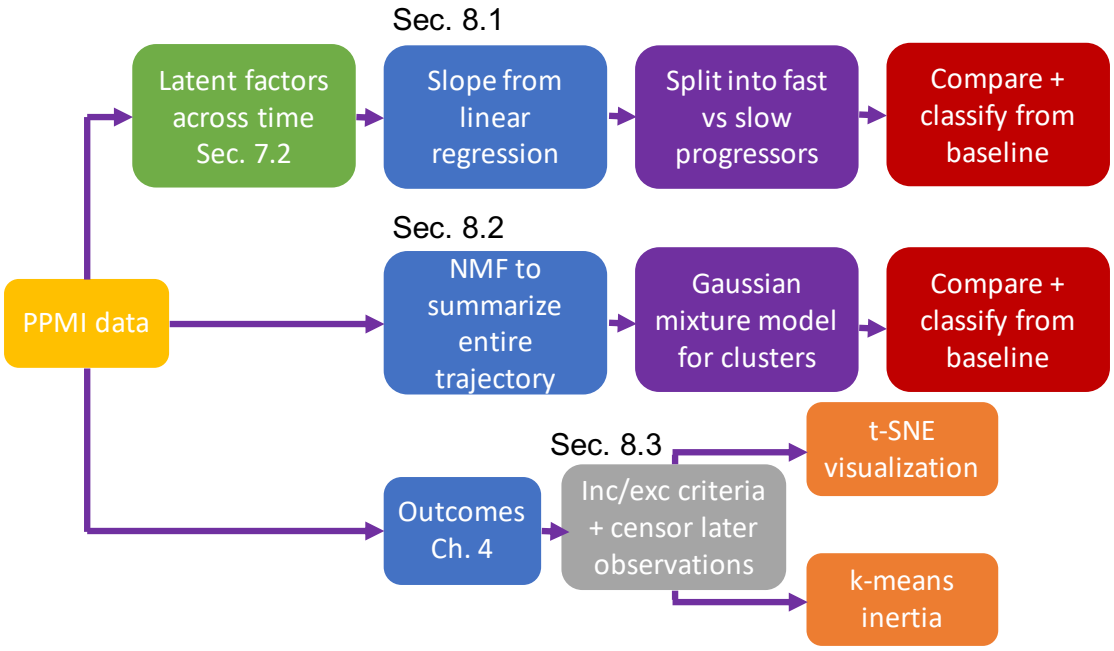

Figure 8-1: **Workflow for three subtyping approaches.**

8.1 Subtyping using latent factor models

Recall that the models in section 7.2 compute latent factors across time for each patient. We start by identifying whether a patient is a slow or fast progressor using these latent factors. A simple method is to fit a linear regression and take the slope. We understand this might not be the best fit, evidenced by the examples in Figure 7-8, but we start with this approach here. With the slopes, we separate the patients into slow and fast progressors at the median rate of the training set. Then, we compare how the distributions of baseline features differ between the two clusters and train a logistic regression predicting cluster membership from baseline features using scikit-learn [59]. Baseline features are min-max normalized to 0-1 using the training range so coefficients are comparable. This section is a continuation of the class project with Suchan Vivatsethachai and Liyang (Sophie) Sun.

As seen in Table 8.1, progression appears to be faster among patients who are affected more severely by the disease on the right side of their body. This may be because most people are right-dominant, and as such, the disease is more likely to hinder their daily life. Lower dopamine levels in DaTscan brain imaging, lower levels

of some CSF biomarkers, and depression are associated with faster progression. A motor subtyping system that is well-established in PD research is tremor-dominant (TD) vs postural instability gait-dominant (PIGD) [78]. Refer to section 2.3.1 for more information. We find that the proportion of patients who are PIGD is higher in the faster-progressing group, aligning with the review by Xia et al [91], but the difference is not statistically significant.

Comparing Tables 8.1 and 8.2, the features that have large coefficients in the classifier differ from those with significantly different distributions across the two clusters. The negative sign on baseline MDS-UPDRS III and absence of age contradict established knowledge. Genetic PCA component 0, which captures single nucleotide polymorphisms associated with catecholamine O-methyltransferase (COMT), has a large positive coefficient. COMT is an enzyme that breaks down dopamine and other neurotransmitters and a frequent PD treatment target, so this result seems sensible. Table 8.3 shows that the classifiers perform similarly for all models. The cluster profiles and coefficients of other models are omitted for brevity.

Characteristic	**Slower cluster**	**Faster cluster**	**p-value**
Latent 0 rate	-0.283	1.020	7.943e-41
Latent 1 rate	-0.035	0.753	1.896e-11
Right-dominant	0.325	0.530	5.423e-05
DaTscan ipsilateral putamen	1.017	0.900	0.003
DaTscan ipsilateral caudate	2.219	2.068	0.013
DaTscan contralateral putamen	0.734	0.659	0.007
DaTscan contralateral caudate	1.909	1.765	0.013
CSF pTau to tTau	0.080	0.073	0.012
CSF pTau log	2.224	1.664	0.013
CSF alpha-synuclein log	7.270	7.180	0.038
Epworth sleep	0.194	0.110	0.025
GDS depressed	0.110	0.188	0.036
TD	0.827	0.823	0.919
PIGD	0.042	0.083	0.104

Table 8.1: **Cluster profiles** at baseline from linear factor analysis model. Means are shown in the two middle columns. p-value is from Welch's t-test. All features with p-value less than 0.05 are shown. Patients who are neither TD nor PIGD are indeterminate.

Feature	Coeff	Feature	Coeff
Right-dominant	1.152	DaTscan ipsilateral putamen	-1.183
GDS depressed	0.988	DaTscan contralateral putamen	-1.066
HVLT retention	0.655	CSF pTau (log-scaled)	-1.043
SCOPA-AUT	0.640	Height (cm)	-0.998
# years education	0.571	MDS-UPDRS III	-0.948
PIGD	0.530	Epworth sleep	-0.883
Right-handed	0.420	Letter-number sequencing	-0.496
Genetic PCA comp 0	0.386	RBD sleep disorder	-0.280
Male	0.297	Symbol-digits modality	-0.275

Table 8.2: Top 9 largest and smallest **coefficients in logistic regression** predicting cluster membership from baseline features. Clusters are determined using the linear factor analysis model.

Method	AUROC	Accuracy	Precision	Recall
Linear FA	0.655 (0.067)	0.599 (0.059)	0.598 (0.063)	0.582 (0.067)
VAE	0.637 (0.067)	0.614 (0.082)	0.619 (0.104)	0.617 (0.105)
Aging CS	0.596 (0.083)	0.452 (0.138)	0.712 (0.361)	0.318 (0.242)
Aging LON	0.656 (0.053)	0.538 (0.057)	0.861 (0.054)	0.466 (0.086)
Lin ord CS	0.550 (0.054)	0.521 (0.032)	0.513 (0.081)	0.544 (0.087)
Nonlin ord CS	0.547 (0.014)	0.527 (0.045)	0.531 (0.036)	0.510 (0.041)
Nonlin ord LON	0.517 (0.031)	0.522 (0.050)	0.531 (0.093)	0.553 (0.057)

Table 8.3: **Evaluation of cluster prediction**. Mean from 5-fold cross-evaluation is shown, followed by standard deviation in parentheses. Precision and recall are for cluster 1. Refer to Table 7.2 for what each method is.

8.2 Non-negative matrix factorization on trajectories

The previous approach starts by finding a latent representation for each timepoint and fitting a linear regression to this latent representation across time to obtain the slope that is used for classifying rate of progression. One potential drawback is that when a linear regression is a poor fit, this slope is a poor representation of the patient. In this alternative, we find a few latent variables to represent the entire trajectory instead. The method we use is called **non-negative matrix factorization (NMF)** [39]. In NMF, the data X is an $n \times d$ matrix, where n is the number of samples and d is the number of features. Then, we choose the number of components h. The objective is to learn the W, an $n \times h$ matrix, and H, an $h \times d$ matrix, that minimize $X - WH$. W holds the h-component representation of each sample, and H is the mapping between the components and the features.

This work closely follows Faghri et al [16] and builds on their open source code. To outline their approach, first they collapse features from across multiple timepoints into a few factors using NMF. Data from every 6 months for the first 4 years are used. The standard set of assessments we used in previous chapters also form the dataset here. The neurological exam is also included here, while biomarkers and demographics are omitted. Min-max normalization is applied. Then, they use a Gaussian mixture model to perform subtyping on these factors. Next, they predict which cluster a patient belongs in using a logistic regression on the baseline features. **Recursive feature elimination (RFE)** and 5-fold cross-validation are used. Recursive feature elimination starts with the full set of features, removes the least important feature (smallest coefficient magnitude), fits a new model, removes the least important feature again, and continues one-by-one until the desired number of features remains.

We examine the 3 components learned from the first step in Figure 8-2. In particular, we see that the healthy controls and genetic unaffected cohorts are well-separated from the *de novo* and genetic PD cohorts when we look at any pair of components. The prodromal cohort is also somewhat between the two in the left and right figures.

Figure 8-3 shows that the 3 components correspond to approximately the same features at each timepoint. Component 0 is most correlated with cognitive, component 1 with motor, and component 2 with sleep. We chose to stay with the choice of 3 components in Faghri et al [16] since with 2 components, sleep seems to be mixed into the 2 components and with 4, one of the components is not capturing much. We also briefly explored PCA and independent component analysis, but those did not separate the features onto the components very well.

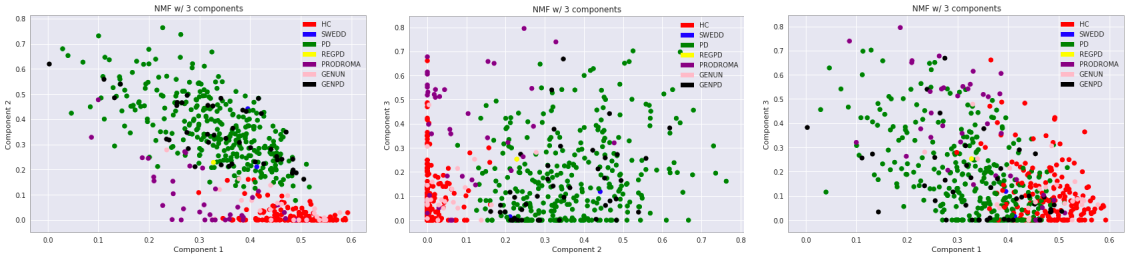

Figure 8-2: **3-component NMF colored by PPMI cohort**. Left: component 0 on x-axis and 1 on y-axis. Middle: component 1 on x-axis and 2 on y-axis. Right: component 0 on x-axis and 2 on y-axis.

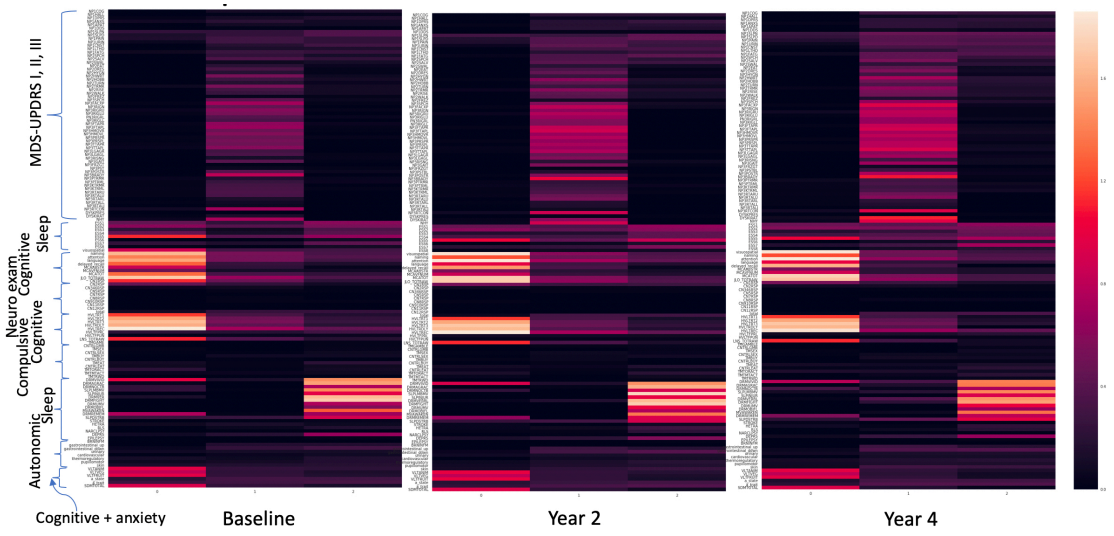

Figure 8-3: **NMF components H at baseline, year 2, and year 4**. Brighter color indicates higher value in H. Columns are the 3 components. Rows are the features.

Next, we examine the subtypes identified in the *de novo* PD cohort using Gaussian mixture models (GMM) and Bayesian GMMs. For GMM, we select the number of clusters that results in the smallest Bayesian information criteria. 3 was the minimum

in the range 1 to 10. The subtypes get progressively farther from healthy controls, as shown in Figure 8-4. Along the "predominantly sleep" component 2, we see the subtypes are clearly ordered with cluster 1 as mild, 2 as moderate, and 0 as severe. Note that the healthy controls primarily fall towards the mild end of the sleep component but some also tend towards the more severe end. To characterize each subtype, we look at the component distributions in Figure 8-6. Using the healthy controls to orient us, we see that higher values for the "predominantly motor" component 1 and "predominantly sleep" component 2 indicate more severe symptoms, while higher values for the "predominantly cognitive" component 0 indicate less severe symptoms. This aligns with how the healthy controls are towards the right in the left and right plots of Figure 8-3.

A Bayesian GMM models the mixture weights and cluster parameters as random variables. It automatically selects the number of clusters, so we simply set a maximum of 10. In this case, 5 clusters are selected, which seems like too many since they are overlapping, as seen in Figure 8-5. The "predominantly sleep" component 2 also seems to order clusters 7, 4, and 5 here. The other clusters seem too small to be meaningful. (The cluster numbers that are omitted had no samples that primarily belonged to them.)

Figure 8-7 plots the totals of questions related to cognitive, motor, and sleep–the 3 categories that were identified. What the subtypes end up capturing is how a patient who has more severe symptoms at baseline is more likely to have severe symptoms at a later timepoint as well, which makes them not as meaningful. Perhaps, if we were to represent the longitudinal data in a way that recognizes two features are the same assessment question taken at two different visits, we might be able to find less obvious features that suggest different rates of progression. As such, predicting which subtype a patient belongs in from baseline is a fairly easy task, evidenced by the high AUROCs in Figure 8-8. We keep the number of features selected by RFE at 54, the number that was chosen by Faghri et al [16]. The features that were picked were mostly a mix of MoCA, MDS-UPDRS parts II and III, and REM sleep questions, with a few HVLT questions, a few QUIP questions, and 1 feature from the neurological

exam included as well.

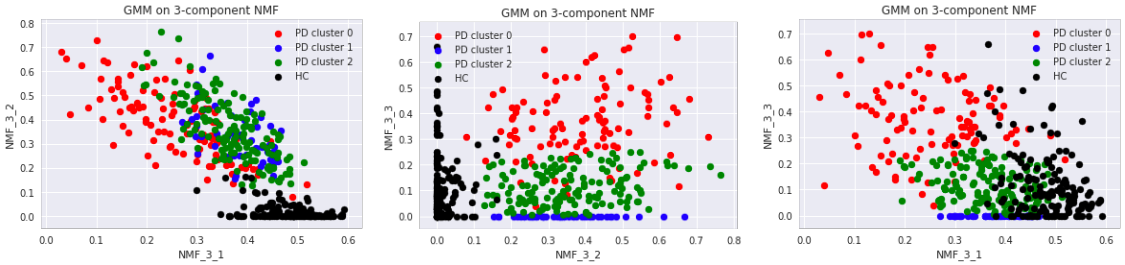

Figure 8-4: **3-component NMF colored by subtypes identified using a GMM**. See Figure 8-2 for description.

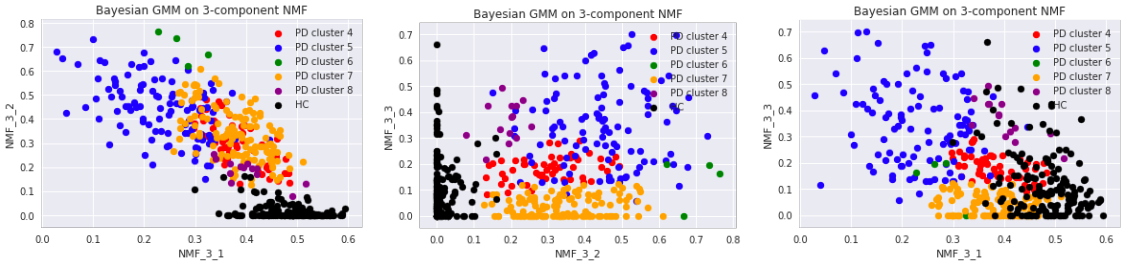

Figure 8-5: **3-component NMF colored by subtypes identified using a Bayesian GMM**. See Figure 8-2 for description.

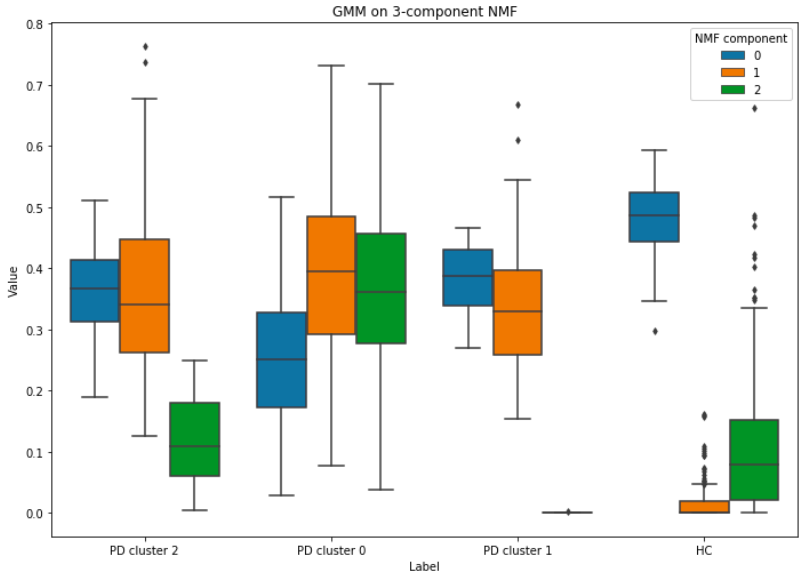

Figure 8-6: **NMF component distributions in the 3 PD clusters identified by a GMM and the HC cohort**. x-axis is PD cluster 2, PD cluster 0, PD cluster 1, and healthy controls.

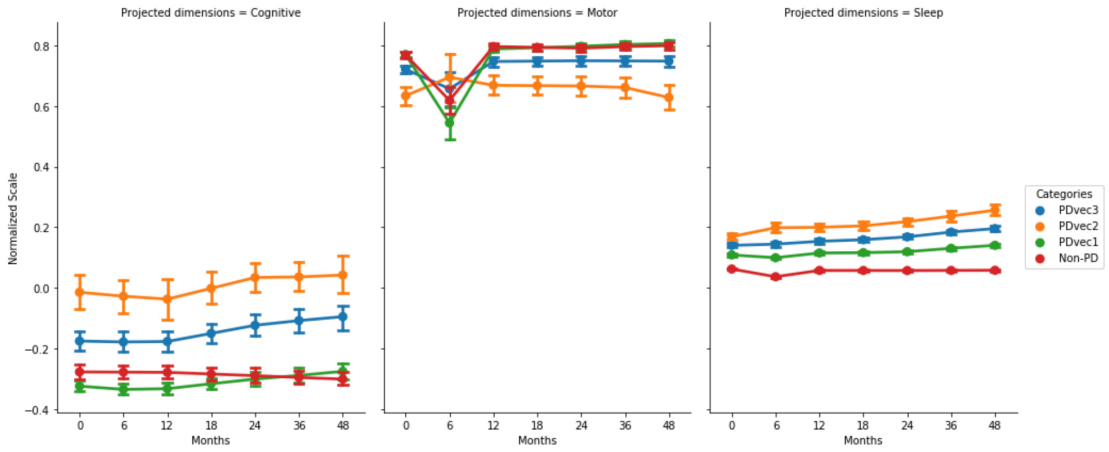

Figure 8-7: **Question totals in cognitive (left), motor (middle), and sleep (right) for the 3 PD subtypes and HC cohort.** In the legend, non-PD refers to healthy controls, PDvec1 refers to PD cluster 0, PDvec2 refers to PD cluster 1, and PDvec3 refers to PD cluster 2.

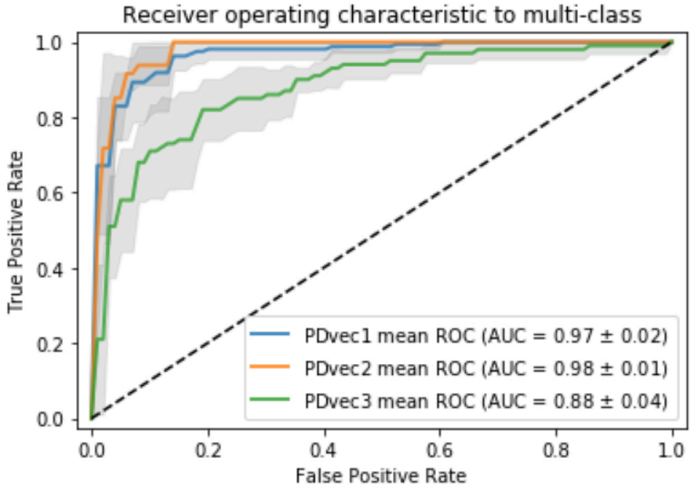

Figure 8-8: **Receiver operating characteristic curve for predicting each subtype from baseline** using a multi-class logistic regression. PDvec1 through PDvec3 refer to clusters 0 through 2. Shaded regions are the standard deviation in the true positive rate at that false positive rate.

8.3 K-means clustering on outcomes

Another way to consider longitudinal subtyping is to use the outcomes we defined in chapter 4 and perform clustering on the outcome times. This assumes that the outcome times are correlated enough across the categories that studying them together would be meaningful. We examine this hypothesis here.

Because the outcomes are censored, we cannot simply plug the observation and censoring times into a clustering model. To handle censoring, we only include patients who are enrolled for at least T years and set any later observation or censoring time to T. This is similar to one of the inclusion-exclusion criteria in chapter 6 except we are free to set a larger T here than the 2- or 3-year trial settings used in chapter 6. We select $T = 4.125$ because that balances the trade-off between maximizing the number of patients included and minimizing the number of patients who are artificially set to T, especially if they were actually observed after T, as seen in Figures 8-9 and 8-10.

We perform **t-distributed stochastic neighbor embedding (t-SNE)** using scikit-learn [59] with 1 or 2 components with the 5 outcome times as input. t-SNE is a nonlinear dimensionality reduction technique designed primarily for visualization since it places similar points nearby and dissimilar points far apart with high probability [41]. The goal of these visualizations is to observe if any clusters arise or if the components are particularly correlated with any of the outcomes. For the latter, we color the visualizations by 4 ranges of outcome times.

As seen in Figure 8-11 for 1-component t-SNE, there are no separate clusters along the x-axis. Only the autonomic outcome shows some correlation with the component, as later times appear more on the right. The other outcomes are much more interspersed. In Figure 8-12 for 2-component t-SNE, we have some groupings. The right side seems to be correlated with later autonomic times. A clump in the upper right seems correlated with earlier psychiatric times, while a clump in the lower left that overlaps in the middle seems correlated with earlier sleep times. The visualization seems to radiate outwards in terms of motor times. Further evaluation is needed to identify clusters.

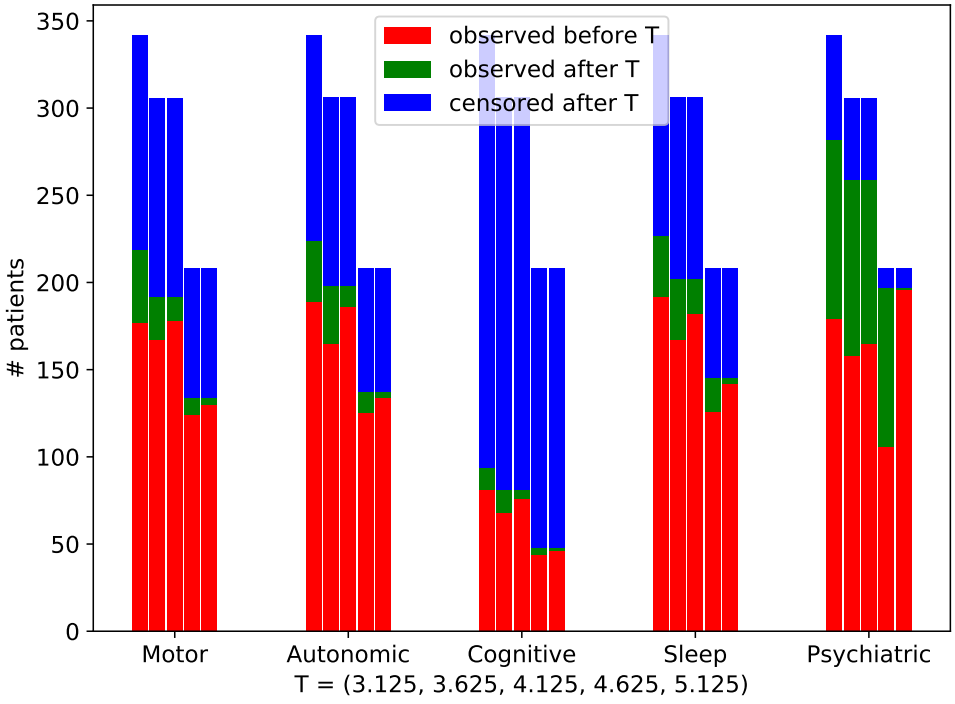

Figure 8-9: **Distributions of observed before T, after T, and censored patients who are included for each choice of T.**

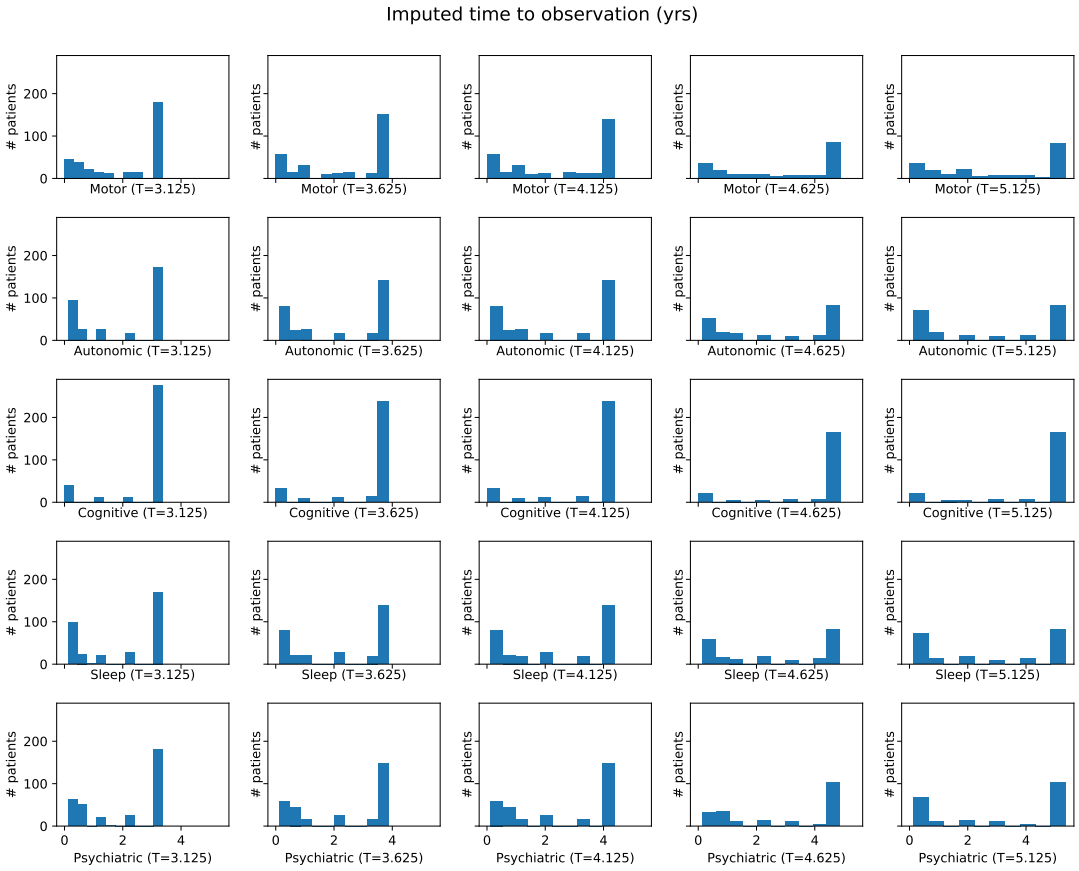

Figure 8-10: **Distributions of outcome observation times after imputing with each choice of T.**

Lastly, we perform k-means clustering on these outcomes. We try to ascertain the number of clusters by measuring inertia. Inertia is the average of the squared distances from each point to its assigned cluster center. As shown in Figure 8-13, both the train and test inertia continue to decrease as we increase the number of clusters. This indicates that it is likely not possible to form a small number of clusters in the data. If there are too many clusters, then it is no longer clinically useful.

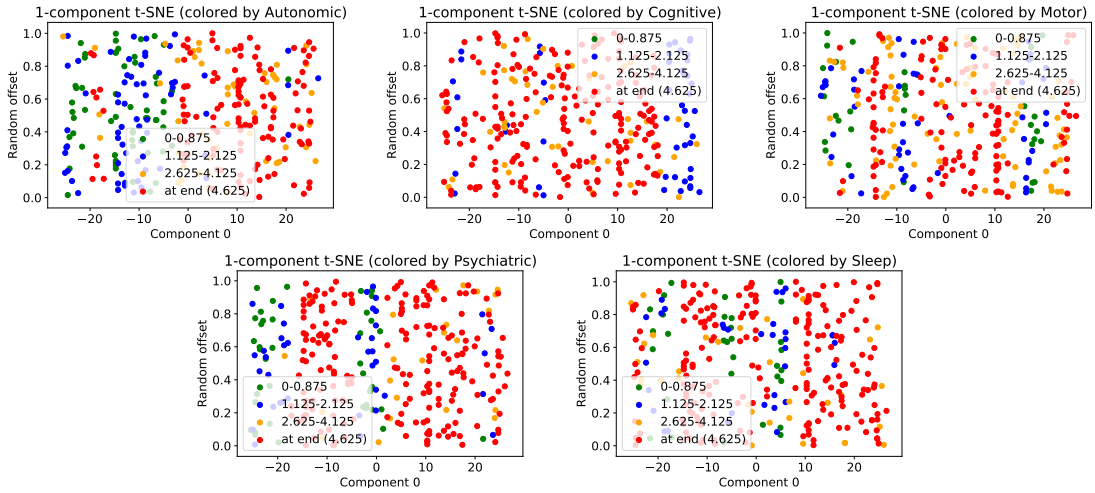

Figure 8-11: **1-component t-SNE on the outcomes** colored by each outcome. x-axis is the component. y-axis is a random offset so the points are not on top of each other on a line.

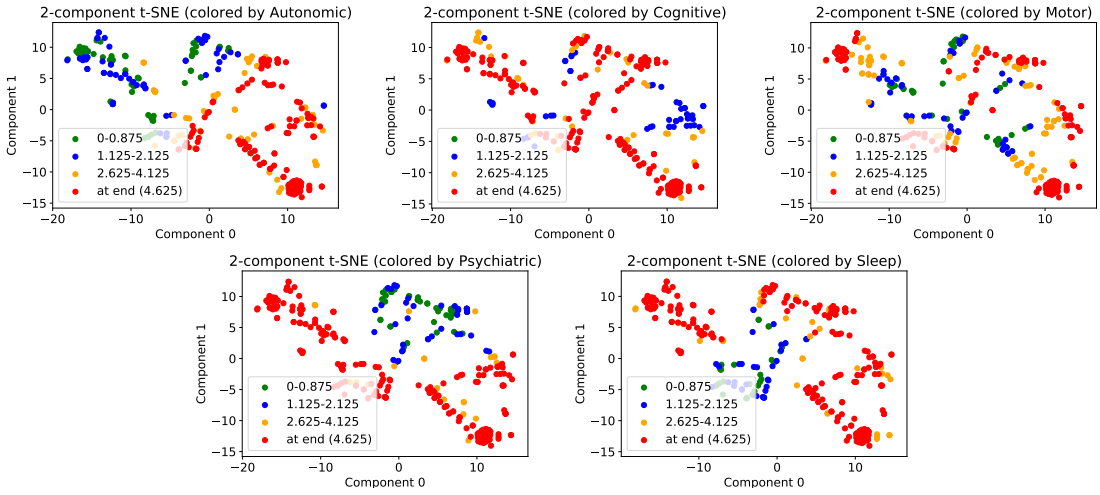

Figure 8-12: **2-component t-SNE on the outcomes** colored by each outcome.

An alternative to imputing censored times is to model the likelihood under a

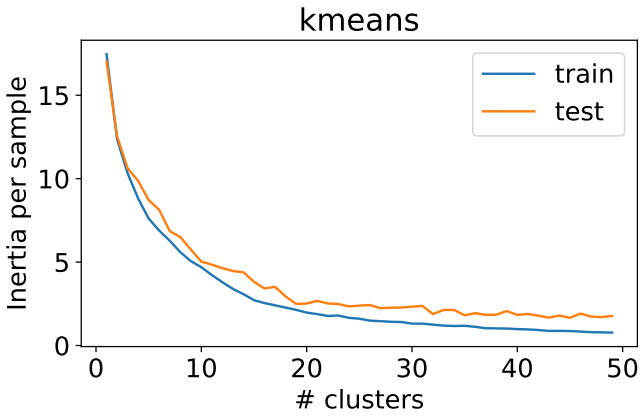

Figure 8-13: **Inertia from k-means clustering on outcomes**.

distribution. Then, the likelihood of an observation occurring during an interval around a visit and the likelihood of an outcome occurring after the censoring time are both well-defined using the cumulative distribution function. However, because the distribution of outcome times in each cluster is unknown, this is likely very sensitive to assumptions when we parameterize the distributions.

8.4 Discussion

In this chapter, we looked at three approaches for machine learning-driven subtyping of Parkinson's disease: 1) using the rate of change in a latent factor, 2) using several components that summarize the entire trajectory, and 3) using outcome times. For 1), we learn how to classify the patients into fast and slow progressors using baseline data. For example, we found that right-dominant disease, lower imaging and CSF features, fewer sleep problems, and depression tend to be associated with faster motor progression. For 2), we profile the clusters in terms of the components and assessments across time. The main conclusion we drew here was that the baseline scores themselves were the most predictive of scores later on. For 3), we believe more technical work may be needed to handle censoring before any conclusions can be reached.

For the first two approaches here, because subtyping is done on latent factors that summarize longitudinal trajectories, it is imperative to first verify that the latent

factors are meaningful by discussing with clinicians. Interpreting the representations, looking for correlations with observed profiles, and examining individual patients are good starting points for this discussion. We hope this chapter opens the door to more subtyping approaches. We particularly focused on the question of how to capture heterogeneity from multiple longitudinal timepoints. We also recommend capturing treatment response patterns in representations for subtyping.

Chapter 9

Discussion

The intersection of machine learning and healthcare has been a burgeoning field. In particular, using machine learning to study the progression of chronic illnesses can shed light on patterns that are predictive of future prognosis, subtypes of patients, and long-term effects of potentially disease-modifying treatments. For this thesis, we focus on studying the progression of Parkinson's disease. From a clinical perspective, there is still much that is unknown about Parkinson's disease. We believe that machine learning can help address some of these questions and make a difference in clinical practice. As applying machine learning to studying Parkinson's disease is relatively new, we hope this thesis opens the door to some promising directions.

To recap, we provide an overview and a machine learning-friendly format of the PPMI dataset in chapter 2. We hope this will make it easier for future machine learning researchers to quickly pick up domain knowledge related to Parkinson's disease, especially if they are not as fortunate as us to have such helpful clinical collaborators. Chapter 3 provides a quick summary of state-of-the-art research related to Parkinson's disease and disease progression modeling in general. Chapters 4 to 6 are the main contribution of this thesis: crafting survival outcomes that represent symptom severity, using these outcomes to predict progression from baseline, and measuring how these outcomes could make clinical trials more efficient. Lastly, we hope to spark more discussion on modeling patient trajectories and subtyping with longitudinal data through our initial forays in chapters 7 and 8.

To conclude in this chapter, we leave the reader with some final thoughts on open clinical questions and how we think machine learning can shed light on these questions. In particular, we call for more attention on the unmet need for better translations from machine learning predictions to clinical answers.

The standard approach for answering the question "How will the disease advance in a few years for a patient?" is to predict changes in quantitative assessments. However, telling a patient their assessment score will increase by 10 points is hardly meaningful. For one, the assessments are very heterogeneous. A 10-point increase that solely comes from a sharp increase in tremor symptoms for example is very different from a 10-point increase that is mostly noise scattered across various questions. By requiring 2-visit persistence and separating the questions by feature category, we are able to reduce the effect of noise in our outcomes while still allowing for heterogeneity. However, a patient would prefer to know what a 10-point increase or outcome corresponds to in their daily life. For example, a patient is more likely to be concerned with when their disease will hinder daily activities or make them dependent on a caregiver. Although there is a scale for daily activities, this scale captures little variation in the first few years of Parkinson's disease.

Measuring treatment effect is also an important question for both clinical decisions and design of clinical trials for identifying disease-modifying therapies. The questions "How will a particular treatment affect a patient?" and "What is the best longitudinal treatment strategy?" are also currently measured by quantitative assessments. Again, a patient would be more interested in how the treatment affects their daily life both in the short-term and long-term. A question would be how long the treatment could delay impairment or dependency. Other facets of treatments are side effects and fluctuations. MDS-UPDRS part IV was designed to measure those for dopaminergic medications. More work on predicting MDS-UPDRS part IV and a more diverse array of such assessments would also be relevant.

Most of the tasks we identified above are supervised learning tasks: given a set of labels Y and a set of covariates X, learn the function $y = f(x)$. Machine learning also has the potential to make unsupervised discoveries: given a set of covariates X and no

labels Y, what patterns can we find? The two most common tasks are clustering and latent factor analysis, corresponding to characterizing patient subtypes and learning a "disease age." A challenge in unsupervised learning in general is confirmation bias: it is easy to confirm if what we discover aligns with what is already known clinically and miss something novel. To go beyond the limits of this bias, we can guide unsupervised methods towards unanswered clinical questions and be open to alternative hypotheses and unformulated questions. A starting point could be identifying ways to evaluate whether future discoveries from these unsupervised explorations are meaningful.

A question that we brought up when introducing the PPMI dataset was that of studying prodromal (pre-diagnosis) progression. Characterizing patients in this phase has garnered a lot of clinical interest, as seen in [44, 57, 63]. Since most neuronal loss predates diagnosis, intervening during the prodromal phase has the most potential for modifying the disease trajectory. We mentioned above that some scales, in particular those on daily activities, fail to capture variation in early-stage PD patients. This holds even more so for prodromal progression. A first step towards this goal is designing assessments that are tailored to prodromal symptoms. One could imagine zooming in on the intermediate stages between biologically normal and diagnosis. A possibility is to build a new assessment by first finding factors of variation and then selecting the questions that most succinctly capture these factors.

Because of how the prodromal criteria are designed, some clinicians hypothesize that the prodromal cohort may only include a subset of future PD patients. Other patients go undetected until diagnosis. One way to start identifying this missing set of pre-diagnosis patients is to identify if they correspond to particular subtypes of PD patients. Once these PD patients can be characterized, additional data can be collected and methods can be developed to augment and better understand the prodromal population.

In conclusion, characterizing and predicting disease progression is an interesting but challenging question. Rather than setting up simple prediction targets, researchers must take the time to consider what patients and clinicians want to know about the disease, especially when making decisions that affect daily life. With this

focus in mind, we hope that machine learning researchers and clinicians can come together to push the forefront of clinical practice.

Appendix A

Additional tables and figures

A.1 Treatment groupings from chapter 2

Additional information about treatments in PPMI are shown here.

#	Grouping	Class keywords	PD freq	HC freq
1	Dopamine replacement	'DOPAMINE REPLACEMENT'	92.9%	2.0%
2	Muscle	'ANTICONVULSANT', 'ANTITREMOR', 'ANTISPASMODIC', 'MUSCLE RELAXER'	20.3%	9.2%
3	Pain	'ANESTHETIC', 'ANALGESIC'	19.1%	1.5%
4	Urinary	'BLADDER CONTROL'	10.9%	1.5%
5	Anxiety	'ANXIOLYTIC'	27.4%	7.1%
6	Depression	'ANTIDEPRESSANT'	38.5%	20.9%
7	Other psychiatric	'ANTIPSYCHOTIC', 'MOOD STABILIZER'	7.3%	2.0%
8	Cognitive	'COGNITIVE ENHANCER'	7.8%	0.0%
9	Sleep	'SLEEP AID'	15.6%	12.2%
10	Digestive	'ANTIEMETIC', 'ANTIDIARRHEAL', 'DIGESTIVE AID', 'ANTACID'	52.5%	34.2%
11	Cardiovascular	'ANTIHYPERTENSIVE', 'ANTIARRHYTHMIC', 'ANTIHYPOTENSIVE', 'ANTICOAGULANT/BLOOD THINNER', 'VASODILATOR', 'ANTIHYPERLIPIDEMIC', 'ANTIHYPERGLYCEMIC'	54.6%	54.6%
12	Anti-inflammatory	'ANTIINFLAMMATORY', 'NSAID'	6.48%	10.2%
13	Immune	'IMMUNOSUPPRESSANT', 'ANTIHISTAMINE', 'ANTIVIRAL', 'ANTIBIOTIC', 'ANTIFUNGAL', 'ANTICANCER', 'VACCINE', 'ADRENALCORTICAL REPLACEMENT'	52.2%	37.8%
14	Respiratory	'DECONGESTANT', 'MUCOLYTIC', 'BRONCHODILATOR'	9.9%	10.2%
15	Thyroid	'THYROID', 'ANTITHYROID AGENT', 'THYROID HORMONE'	27.7%	18.4%
16	Supplement	'SUPPLEMENT', 'PD SUPPLEMENT', 'BONE/JOINT HEALTH'	76.4%	69.9%
17	Eye	'OPTHALAMIC', 'ANTIGLAUCOMA'	9.2%	5.6%
18	Other	'CONTRACEPTIVE', 'ANTI BPH', 'NEUROTOXIN', 'DERMATOLOGIC', 'STIMULANT', 'HORMONE REPLACEMENT', 'URIC ACID REDUCER', etc.	43.5%	42.3%

Table A.1: **Groupings of drug classes** and percentage of subjects in PD and HC cohorts who use them at least once.

A.2 Additional coefficient tables from chapter 5

The 2 best-performing hybrid outcome survival analysis models had a large number of significant coefficients. To keep the main text concise, the coefficient tables are shown here.

Feature	Coefficient
MoCA clock contour	0.2368 (0.0998)
BJLO3	0.1907 (0.0613)
contralateral caudate	0.1340 (0.0503)
ipsilateral caudate	0.1337 (0.0445)
MoCA tap foot on A	0.1250 (0.0380)
ipsilateral putamen	0.1162 (0.0233)
contralateral putamen	0.1075 (0.0381)
HVLT recognition	0.1007 (0.0340)
HVLT retention 1	0.0859 (0.0410)
HVLT retention 2	0.0839 (0.0393)
Semantic fluency (animals)	0.0684 (0.0191)
HVLT delayed recall	0.0613 (0.0167)
UPSIT (smell)	0.0582 (0.0053)
STAI 27 calm/cool/collected	0.0495 (0.0231)
Supine diastolic BP	0.0468 (0.0171)
STAI 34 make decisions easily	0.0458 (0.0102)
Semantic fluency (fruits)	0.0418 (0.0127)
Standing systolic BP	0.0408 (0.0164)
HVLT retention 3	0.0380 (0.0060)
Standing diastolic BP	0.0378 (0.0168)
STAI 23 satisfied with myself	0.0366 (0.0186)
BJLO7	0.0328 (0.0135)
STAI 20 pleasant	0.0321 (0.0152)
LNS 2B	0.0320 (0.0152)
LNS 3A	0.0318 (0.0136)
GDS satisfied with life	0.0283 (0.0103)
MoCA recall word 5	0.0268 (0.0013)
STAI 19 steady	0.0268 (0.0090)
MoCA recall word 4	0.0263 (0.0118)
STAI 21 pleasant	0.0253 (0.0040)
MoCA recall word 2	0.0248 (0.0073)
MDS 3.17 rest tremor amplitude left upper	0.0245 (0.0066)
White	0.0233 (0.0084)
STAI 16 content	0.0232 (0.0110)
STAI 11 self-confident	0.0224 (0.0078)
GDS full of energy	0.0207 (0.0091)
STAI 10 comfortable	0.0204 (0.0034)
STAI 36 content	0.0194 (0.0062)

Table A.2: **Coefficients for hybrid outcome** from Weibull model with question + treatment + imaging + CSF + expanded + genetic set of covariates (part 1 of 3). This model had the highest CI. Standard deviation across 4 folds shown in parentheses. Of the 266 features, the 89 shown in these tables had significant coefficients.

Feature	Coefficient
SCOPA male unable ejaculate	-0.0303 (0.0078)
MDS 3.7 toe tapping right	-0.0303 (0.0089)
STAI 22 nervous + restless	-0.0311 (0.0106)
Digestive aid	-0.0344 (0.0132)
injured self/partner in sleep (REM 5)	-0.0346 (0.0120)
SCOPA strain pass stool	-0.0351 (0.0162)
MDS 2.6 hygiene	-0.0380 (0.0171)
MDS 3.3 rigidity right upper	-0.0389 (0.0178)
MDS 3.3 rigidity right lower	-0.0404 (0.0198)
speak in dreams (REM 6.1)	-0.0419 (0.0144)
SCOPA male impotent	-0.0424 (0.0185)
Age	-0.0432 (0.0071)
STAI 33 secure	0.0179 (0.0021)
STAI 30 happy	0.0170 (0.0080)
MoCA copy cube	0.0163 (0.0034)
MoCA connect 1 to A to 2...	0.0131 (0.0063)
MoCA clock hands	0.0126 (0.0057)
BJLO21	0.0105 (0.0008)
MoCA recall word 1	0.0101 (0.0048)
pTau (log)	0.0068 (0.0032)
vivid dreams (REM 1)	-0.0071 (0.0027)
QUIP too much PD med use	-0.0134 (0.0051)
MDS1 depression	-0.0147 (0.0070)
movements wake patient (REM 7)	-0.0160 (0.0039)
disturbed sleep (REM 9)	-0.0177 (0.0081)
arms/legs move in sleep (REM 4)	-0.0191 (0.0064)
MDS 3.8 leg agility left	-0.0194 (0.0048)
Hoehn & Yahr	-0.0203 (0.0048)
STAI 29 worry too much over trivial	-0.0205 (0.0081)
SCOPA intolerant of cold	-0.0227 (0.0031)
sleep on car for 1hr (Epworth 4)	-0.0279 (0.0131)
MDS 3.4 finger tapping right	-0.0285 (0.0119)
SCOPA urine med	-0.0286 (0.0102)

Table A.3: **Coefficients for hybrid outcome** from Weibull model with question + treatment + imaging + CSF + expanded + genetic set of covariates (part 2 of 3). See previous table for description.

Feature	Coefficient
MDS1 light-headed	-0.0436 (0.0169)
sudden move in dreams (REM 6.2)	-0.0438 (0.0216)
MDS 3.8 leg agility right	-0.0475 (0.0170)
tTau-to-Abeta ratio	-0.0506 (0.0155)
MDS 3.2 facial expression	-0.0570 (0.0191)
QUIP too much buying	-0.0596 (0.0274)
SCOPA urine weak	-0.0599 (0.0271)
SCOPA lightheaded upon standing	-0.0613 (0.0253)
MDS 3.14 global spontaneity of movement/body bradykinesia	-0.0616 (0.0172)
MDS1 daytime sleep	-0.0618 (0.0192)
SCOPA involuntary urine	-0.0656 (0.0199)
MDS 3.13 posture	-0.0693 (0.0151)
MDS 3.1 speech	-0.0722 (0.0335)
SCOPA difficult retain urine	-0.0757 (0.0280)
SCOPA drool	-0.0803 (0.0384)
MDS 2.1 speech	-0.0885 (0.0437)
MDS 2.11 getting out of bed/car/deep chair	-0.0905 (0.0331)
SCOPA lightheaded after stand some time	-0.1001 (0.0404)

Table A.4: **Coefficients for hybrid outcome** from Weibull model with question + treatment + imaging + CSF + expanded + genetic set of covariates (part 3 of 3). See first table for description.

Feature	Coefficient
MoCA clock contour	0.2904 (0.0725)
ipsilateral caudate	0.2025 (0.0490)
contralateral caudate	0.1834 (0.0442)
ipsilateral putamen	0.1816 (0.0600)
contralateral putamen	0.1762 (0.0442)
MoCA tap foot on A	0.1646 (0.0511)
LNS 2C	0.1503 (0.0521)
Semantic fluency (vegetables)	0.1397 (0.0484)
BJLO3	0.1349 (0.0299)
HVLT recognition	0.1332 (0.0210)
UPSIT (smell)	0.1158 (0.0277)
Semantic fluency (animals)	0.1090 (0.0471)
HVLT retention 1	0.0958 (0.0190)
MoCA subtract 7 series	0.0947 (0.0304)
HVLT retention 2	0.0927 (0.0143)
MoCA repeat numbers backward	0.0919 (0.0302)
BJLO9	0.0845 (0.0279)
SCOPA fainted past 6 months	0.0755 (0.0373)
STAI 34 make decisions easily	0.0711 (0.0295)
STAI 27 calm/cool/collected	0.0699 (0.0151)
HVLT delayed recall	0.0663 (0.0049)
STAI 26 rested	0.0595 (0.0202)
BJLO23	0.0573 (0.0177)
STAI 23 satisfied with myself	0.0516 (0.0194)
BJLO25	0.0469 (0.0184)
BJLO7	0.0465 (0.0079)
LNS 4A	0.0454 (0.0179)
Standing diastolic BP	0.0426 (0.0167)
LNS 3A	0.0418 (0.0152)
HVLT retention 3	0.0413 (0.0134)
MoCA recall word 4	0.0382 (0.0124)
STAI 19 steady	0.0363 (0.0123)
STAI 20 pleasant	0.0355 (0.0135)
MoCA recall word 2	0.0298 (0.0145)
MDS 3.17 rest tremor amplitude left upper	0.0294 (0.0123)
GDS most better off than self	0.0287 (0.0104)
MoCA copy cube	0.0265 (0.0060)
STAI 21 pleasant	0.0257 (0.0101)
BJLO19	0.0241 (0.0121)

Table A.5: **Coefficients for hybrid outcome** from Weibull model with question + treatment + imaging + CSF + expanded set of covariates (part 1 of 3). This model had the lowest MAE. Standard deviation across 4 folds shown in parentheses. Of the 265 features, the 122 shown in these tables had significant coefficients.

Feature	Coefficient
STAI 33 secure	0.0232 (0.0055)
STAI 16 content	0.0214 (0.0090)
GDS full of energy	0.0191 (0.0075)
STAI 36 content	0.0188 (0.0095)
STAI 10 comfortable	0.0163 (0.0082)
BJLO21	0.0137 (0.0030)
STAI 8 satisfied	-0.0180 (0.0068)
vivid dreams (REM 1)	-0.0184 (0.0029)
remember dreams (REM 8)	-0.0198 (0.0075)
arms/legs move in sleep (REM 4)	-0.0226 (0.0045)
movements wake patient (REM 7)	-0.0230 (0.0103)
disturbed sleep (REM 9)	-0.0237 (0.0044)
Male	-0.0251 (0.0072)
MDS 3.3 rigidity upper left	-0.0283 (0.0112)
MDS 3.4 finger tapping left	-0.0294 (0.0047)
Hoehn & Yahr	-0.0297 (0.0050)
MDS 3.8 leg agility left	-0.0299 (0.0090)
MDS 3.5 hand movements left	-0.0305 (0.0153)
Gastrointestinal history	-0.0335 (0.0167)
SCOPA male unable ejaculate	-0.0365 (0.0149)
MDS 2.7 handwriting	-0.0374 (0.0144)
SCOPA intolerant of cold	-0.0380 (0.0114)
STAI 32 lack self-confidence	-0.0389 (0.0131)
MDS 3.16 kinetic tremor right hand	-0.0398 (0.0196)
Antidepressant	-0.0405 (0.0148)
SCOPA urine med	-0.0419 (0.0146)
MDS 3.4 finger tapping right	-0.0421 (0.0036)
SCOPA male impotent	-0.0449 (0.0084)
Digestive aid	-0.0453 (0.0162)
sleep on car for 1hr (Epworth 4)	-0.0471 (0.0142)
MDS 3.7 toe tapping right	-0.0522 (0.0168)
injured self/partner in sleep (REM 5)	-0.0525 (0.0143)
MDS 2.6 hygiene	-0.0545 (0.0089)
MDS 3.3 rigidity lower left	-0.0555 (0.0111)
MDS1 pain	-0.0563 (0.0155)
sleep when watch TV (Epworth 2)	-0.0571 (0.0240)
MDS 3.5 hand movements right	-0.0575 (0.0261)
MDS 3.3 rigidity right lower	-0.0578 (0.0112)
SCOPA involuntary stool	-0.0599 (0.0258)
SCOPA intolerant of heat	-0.0599 (0.0269)
MDS 3.8 leg agility right	-0.0607 (0.0086)

Table A.6: **Coefficients for hybrid outcome** from Weibull model with question + treatment + imaging + CSF + expanded set of covariates (part 2 of 3). See previous table for description.

Feature	Coefficient
SCOPA urine again w/in 2hrs	-0.0607 (0.0282)
speak in dreams (REM 6.1)	-0.0610 (0.0147)
HVLT false positive related	-0.0639 (0.0177)
MDS 3.3 rigidity right upper	-0.0647 (0.0284)
aggressive/action dreams (REM 2)	-0.0669 (0.0304)
sudden move in dreams (REM 6.2)	-0.0682 (0.0155)
sleep after lunch (Epworth 7)	-0.0691 (0.0294)
SCOPA urine weak	-0.0733 (0.0212)
STAI 37 unimportant thought bothers me	-0.0739 (0.0319)
MDS 2.4 eating tasks	-0.0760 (0.0347)
MDS1 light-headed	-0.0778 (0.0223)
Age	-0.0800 (0.0289)
sleep when sit in public (Epworth 3)	-0.0822 (0.0230)
MDS 2.9 turning in bed	-0.0860 (0.0370)
SCOPA urine at night	-0.0864 (0.0203)
SCOPA involuntary urine	-0.0874 (0.0275)
things near bed fall in dreams (REM 6.4)	-0.0880 (0.0376)
ipsilateral count density ratio	-0.0907 (0.0306)
MDS 3.14 global spontaneity of movement/body bradykinesia	-0.0937 (0.0202)
MDS1 fatigue	-0.0938 (0.0255)
MDS1 daytime sleep	-0.0940 (0.0462)
MDS 3.2 facial expression	-0.0959 (0.0211)
MDS 2.5 dressing	-0.0991 (0.0289)
MDS 3.1 speech	-0.1014 (0.0190)
MDS1 constipation	-0.1023 (0.0264)
sleep on stopped car (Epworth 8)	-0.1027 (0.0495)
SCOPA lightheaded upon standing	-0.1056 (0.0396)
MDS 3.13 posture	-0.1066 (0.0124)
SCOPA difficult retain urine	-0.1072 (0.0246)
SCOPA can't completely empty bladder	-0.1083 (0.0380)
MDS 2.1 speech	-0.1131 (0.0404)
MDS 2.2 saliva and drooling	-0.1223 (0.0364)
MDS 2.11 getting out of bed/car/deep chair	-0.1243 (0.0168)
SCOPA drool	-0.1302 (0.0338)
SCOPA lightheaded after stand some time	-0.1327 (0.0377)
MDS 2.12 walking and balance	-0.1364 (0.0540)
MDS 2.8 hobbies and other activities	-0.1408 (0.0347)
MDS1 urinary	-0.1472 (0.0458)
SCOPA swallow difficult/choke	-0.1621 (0.0614)
SCOPA food stuck	-0.1741 (0.0513)
MDS 2.13 freezing	-0.1996 (0.0754)
MDS 3.9 arising from chair	-0.2484 (0.0910)

Table A.7: **Coefficients for hybrid outcome** from Weibull model with question + treatment + imaging + CSF + expanded set of covariates (part 3 of 3). See first table for description.

A.3 Regularization tuning from chapter 6

By plotting the evaluation metrics against the regularization parameter settings in Figures A-1 to A-5, we again verify that applying insufficient regularization sometimes leads to overfitting to the train dataset, so regularization is indeed necessary. Applying excessive regularization leads to underfitting, so we have covered a sufficiently large range of regularization coefficients. Note that the C hyperparameter in logistic regression is inverse to the regularization strength. Most of the time, the trends are shared across the different evaluation metrics. The wild fluctuations in response to the Weibull penalizer are a bit troublesome, as that is also when the metrics stop aligning. Not depicted is L1 ratio. Strangely, L2 regularization is favored when there is a large number of features, while L1 regularization is favored when the covariate set is smaller.

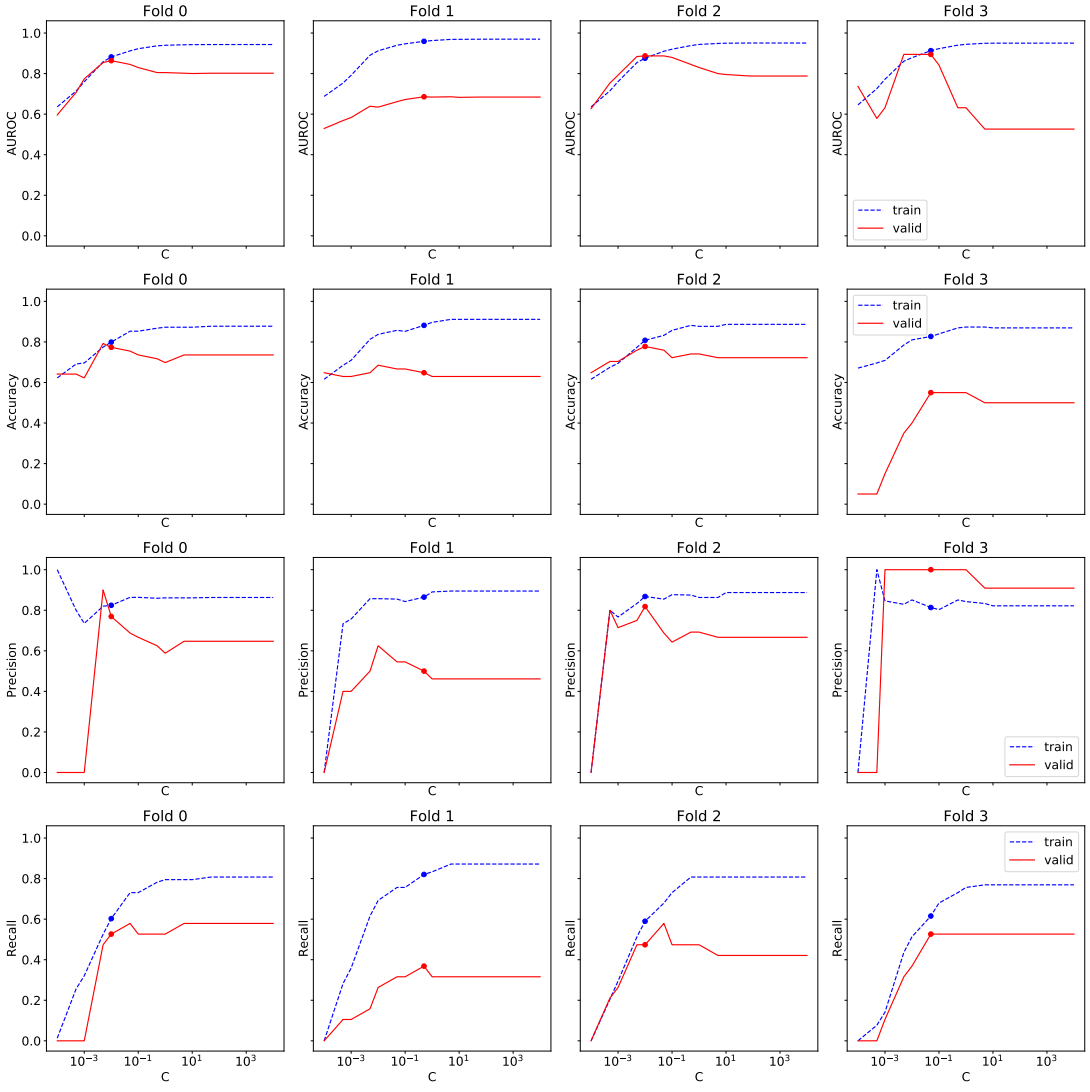

Figure A-1: **Effect of varying C in logistic regression** model of MDS-UPDRS moderate outcome in the 3-year trial setting. Covariate set is Qst + Img + CSF + Exp.

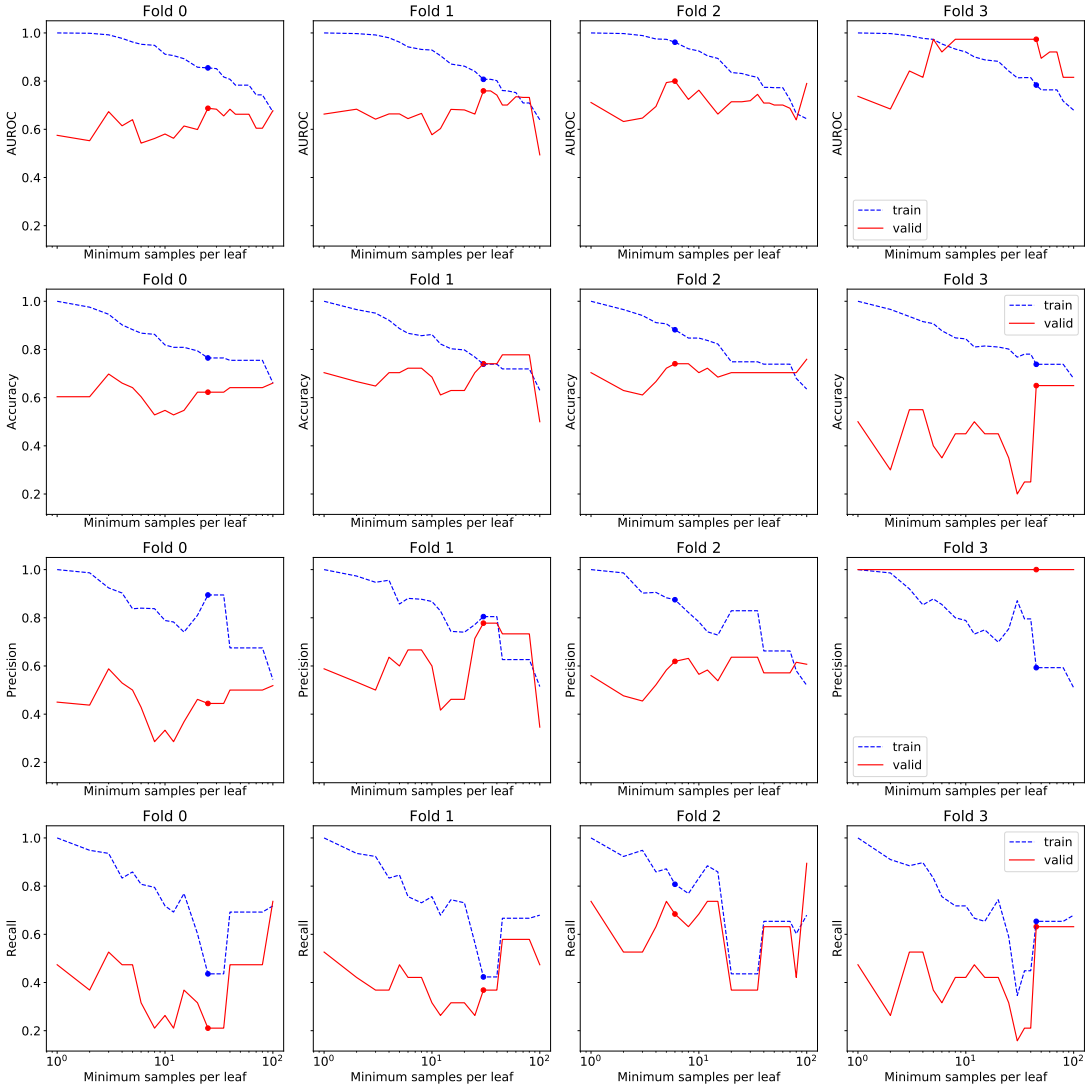

Figure A-2: **Effect of varying minimum number of samples in a leaf in decision tree** model of MDS-UPDRS moderate outcome in the 3-year trial setting. Covariate set is Qst + Img + CSF + Exp.

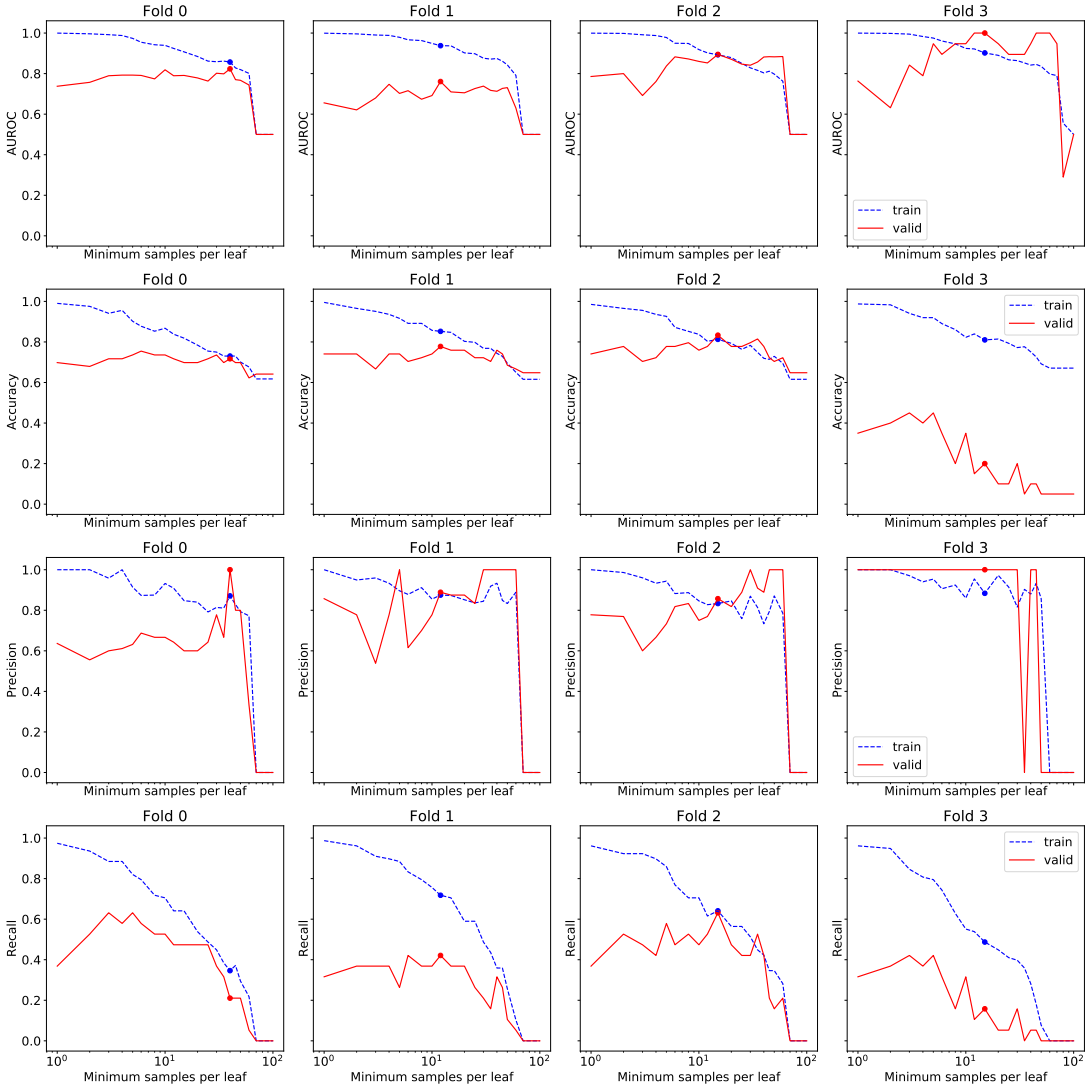

Figure A-3: **Effect of varying minimum number of samples in a leaf in random forest** model of MDS-UPDRS moderate outcome in the 3-year trial setting. Covariate set is Qst + Img + CSF + Exp.

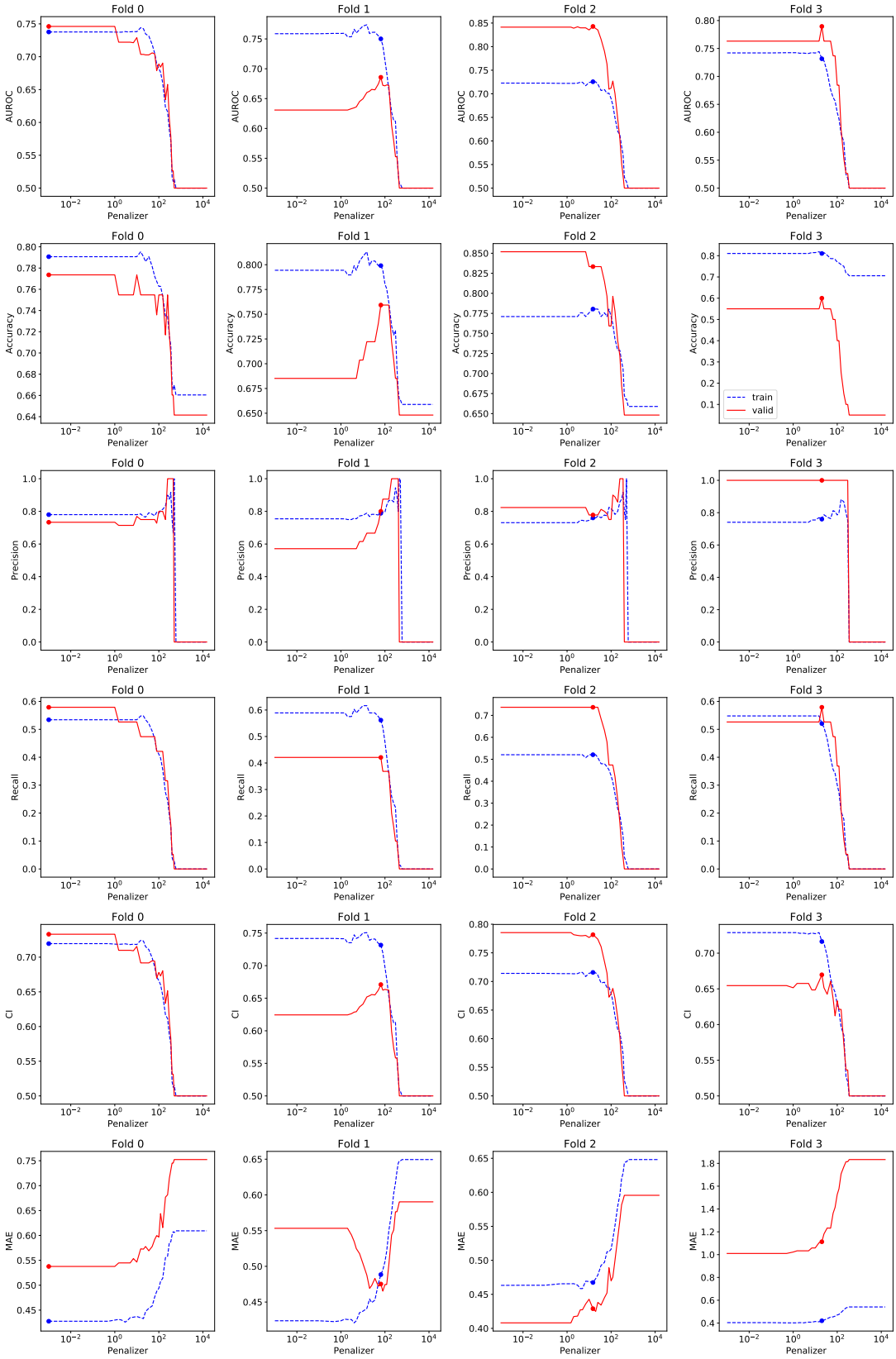

Figure A-4: **Effect of varying penalizer in Cox** model of MDS-UPDRS moderate outcome in the 3-year trial setting. Covariate set is BL.

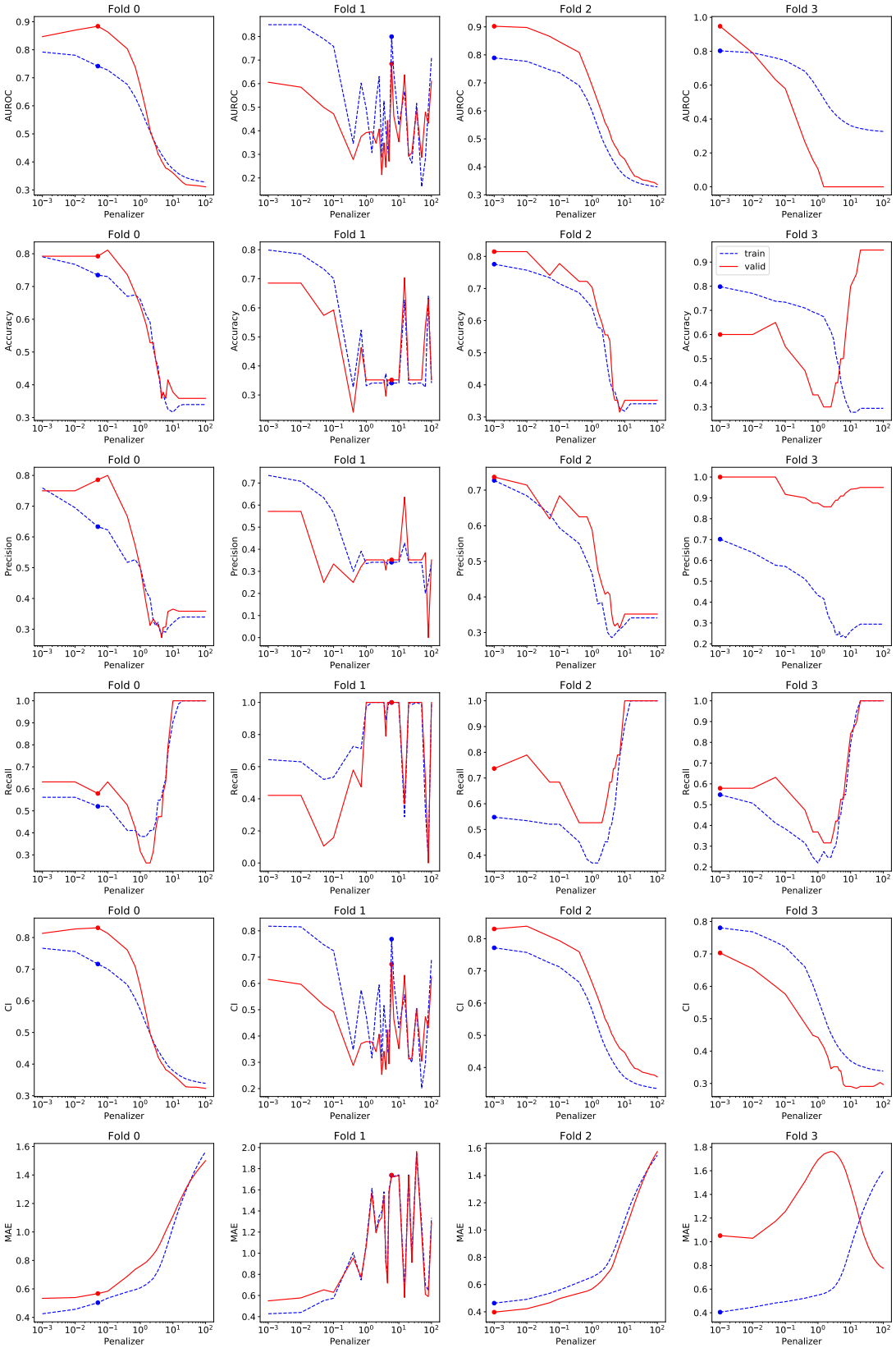

Figure A-5: **Effect of varying penalizer in Weibull** model of MDS-UPDRS moderate outcome in the 3-year trial setting. Covariate set is BL.

A.4 Additional metrics from chapter 6

Some supplementary metrics for the models predicting whether an outcome will be observed during the trial period are shown here.

Model	Accuracy	Recall	CI	MAE
Autonomic 1	0.8269 (0.0000)	0.8100 (0.0520)		
Autonomic 2	0.8125 (0.0083)	0.6400 (0.0000)	0.8273 (0.0037)	0.4165 (0.0039)
Cognitive 1	0.9025 (0.0251)	0.4062 (0.2400)		
Cognitive 2	0.9153 (0.0208)	0.3750 (0.1531)	0.6875 (0.0765)	0.1716 (0.0252)
Psychiatric 1	0.5385 (0.1154)	0.6818 (0.3182)	0.7575 (0.0412)	0.7309 (0.1762)
Psychiatric 2	0.6827 (0.0096)	0.2955 (0.0277)	0.6408 (0.0018)	0.5517 (0.0118)
Motor 1	0.7321 (0.0253)	0.4500 (0.0354)		
Motor 2	0.7411 (0.0155)	0.3500 (0.0000)	0.6424 (0.0070)	0.3790 (0.0075)
Sleep 1	0.6250 (0.1109)	0.7308 (0.1609)	0.7285 (0.0967)	0.5981 (0.1951)
Sleep 2	0.6875 (0.0217)	0.4904 (0.1499)	0.7013 (0.0528)	0.4964 (0.0298)
Hybrid 1	0.8112 (0.0302)	0.5167 (0.1658)		
Hybrid 2	0.8673 (0.0228)	0.6833 (0.0289)	0.8153 (0.0103)	0.2321 (0.0143)
MDS-UPDRS 1	0.8147 (0.0429)	0.4118 (0.1248)		
MDS-UPDRS 2	0.7500 (0.0286)	0.1471 (0.0975)		
MoCA 1	0.8261 (0.0266)	0.5750 (0.1639)		
S & E 1	0.8958 (0.0000)	0.0000 (0.0000)		
S & E 2	0.8958 (0.0147)	0.1500 (0.0866)	0.5698 (0.0410)	0.1406 (0.0172)

Table A.8: **Accuracy, recall, CI, and MAE** for best-performing models for **2-year trial setting**. CI and MAE are present only for survival models. Standard deviation across 4 folds in parentheses. MDS-UPDRS refers to the moderate version of the outcome. S & E stands for Schwab & England.

Model	Accuracy	Recall	CI	MAE
Autonomic 3	0.7806 (0.0088)	0.7212 (0.0500)		
Autonomic 4	0.7041 (0.1005)	0.9135 (0.0687)	0.8947 (0.0000)	0.8250 (0.3327)
Autonomic 5	0.7704 (0.0088)	0.6635 (0.0500)	0.8216 (0.0116)	0.7183 (0.0423)
Cognitive 3	0.8482 (0.0155)	0.2750 (0.1639)		
Cognitive 4	0.8795 (0.0264)	0.325 (0.1479)	0.6669 (0.0759)	0.2215 (0.0677)
Psychiatric 3	0.6633 (0.0338)	0.5100 (0.0768)		
Psychiatric 4	0.5969 (0.0088)	0.3000 (0.0447)	0.6423 (0.0149)	0.9292 (0.0670)
Motor 3	0.5721 (0.1492)	0.7841 (0.2165)	0.7419 (0.0207)	1.0705 (0.5077)
Motor 4	0.7404 (0.0215)	0.4318 (0.0227)	0.6719 (0.0145)	0.5656 (0.0148)
Sleep 3	0.7588 (0.0076)	0.7155 (0.0149)		
Sleep 4	0.8207 (0.0541)	0.6250 (0.1382)		
Hybrid 3	0.8207 (0.0541)	0.6250 (0.1382)		
Hybrid 4	0.7663 (0.0237)	0.4306 (0.0461)		
MDS-UPDRS 3	0.8056 (0.0207)	0.5263 (0.0744)		
MDS-UPDRS 4	0.8241 (0.0404)	0.5526 (0.0873)	0.7506 (0.0400)	0.4234 (0.0262)
MoCA 3	0.6250 (0.1932)	0.7308 (0.1586)	0.7551 (0.0368)	0.7407 (0.6662)
MoCA 4	0.7898 (0.0492)	0.4615 (0.1439)	0.6904 (0.0619)	0.3345 (0.0464)
S & E 3	0.8889 (0.0000)	0.1667 (0.0000)		

Table A.9: **Accuracy, recall, CI, and MAE** for best-performing models for **3-year trial setting**. Refer to Table A.8 for description.

A.5 Feature contributions from chapter 6

We list the coefficients from a logistic regression in Table A.10 and the feature importances as calculated by scikit-learn [59] from a random forest in Table A.11. To understand the random forest better, we visualize the first 5 layers of the decision trees from the same 4 folds of the dataset in Figures A-6 to A-9. Although some of the top features are similar, we note that the decision trees are quite different among the folds, suggesting that having an ensemble as in a random forest might be better. The coefficients for the survival models are similar to those in chapter 5, so they are omitted here.

Feature	Coefficient
MDS3 face	0.2191 (0.1888)
MDS2 daily activities	0.1434 (0.0415)
SCOPA-autonomic	0.1046 (0.0201)
MDS3 right rigidity	0.0651 (0.0618)
STAI (anxiety)	0.0166 (0.0151)
MDS3 left rigidity	0.0094 (0.0070)
caudate asymmetry index	0.0049 (0.0045)
putamen asymmetry index	-0.0114 (0.0085)
UPSIT (smell)	-0.0190 (0.0134)
HVLT discrim recog	-0.1238 (0.1107)

Table A.10: **Coefficients of logistic regression** model of hybrid outcome in the 3-year trial setting. Covariate set is Std + Trt + Img + CSF + Exp. Standard deviation across 4 folds in parentheses. Only the features that have coefficient means more than 1 standard deviation away from 0 are shown.

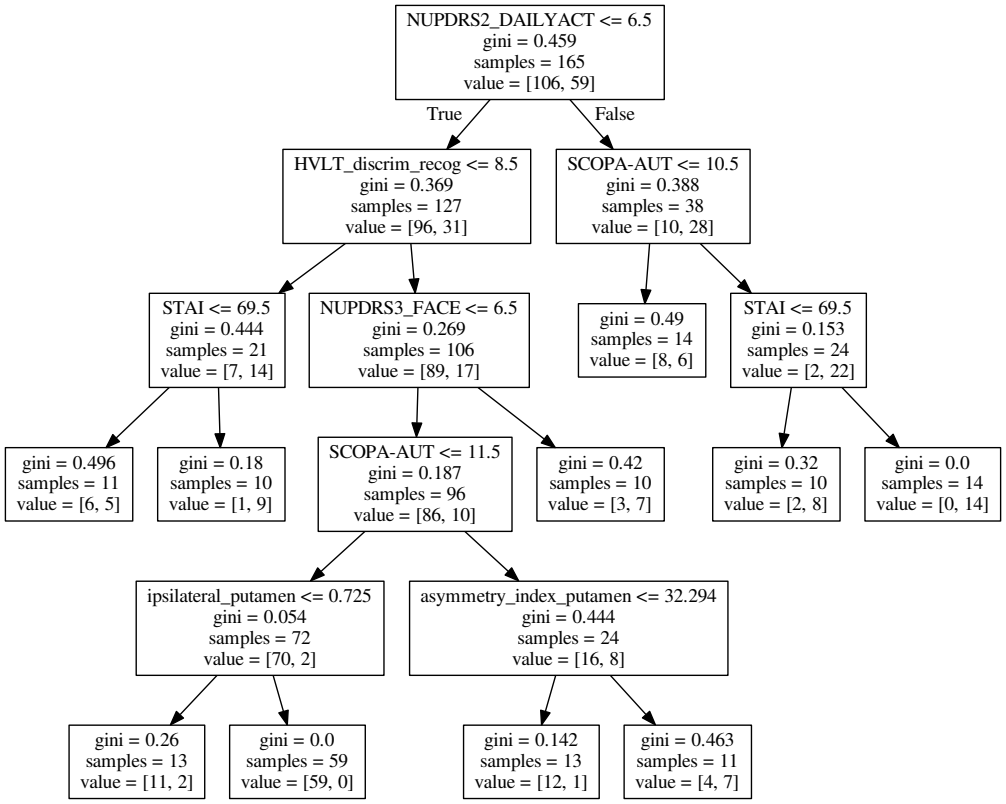

Figure A-6: **Decision tree** model of hybrid outcome in the 3-year trial setting (fold 0). Covariate set is Std + Trt + Img + CSF + Exp. Only first 5 layers are shown.

Feature	Coefficient
MDS3 face	0.1290 (0.0537)
MDS2 daily activities	0.1077 (0.0276)
ipsilateral caudate	0.0933 (0.0308)
MDS3 gait	0.0564 (0.0374)
SCOPA-autonomic	0.0425 (0.0204)
ipsilateral putamen	0.0381 (0.0093)
Semantic fluency	0.0326 (0.0225)
contralateral caudate	0.0323 (0.0236)
MDS1 apathy	0.0252 (0.0237)
HVLT retention	0.0237 (0.0115)
UPSIT (smell)	0.0206 (0.0156)
caudate asymmetry index	0.0190 (0.0096)
HVLT immed recall	0.0176 (0.0135)
ipsilateral count density ratio	0.0172 (0.0114)
MDS1 urinary	0.0172 (0.0126)
pTau-to-Abeta ratio	0.0156 (0.0140)
tTau (log)	0.0154 (0.0069)
Standing systolic BP	0.0116 (0.0083)
Male	0.0102 (0.0095)
Abeta (log)	0.0102 (0.0074)
contralateral count density ratio	0.0099 (0.0073)
pTau-to-tTau ratio	0.0092 (0.0071)
contralateral putamen	0.0092 (0.0072)
Alpha-synuclein (log)	0.0077 (0.0066)

Table A.11: **Feature importances in random forest** model of hybrid outcome in the 3-year trial setting. Covariate set is Std + Trt + Img + CSF + Exp. Standard deviation across 4 folds in parentheses. Only the features that have coefficient means more than 1 standard deviation away from 0 are shown.

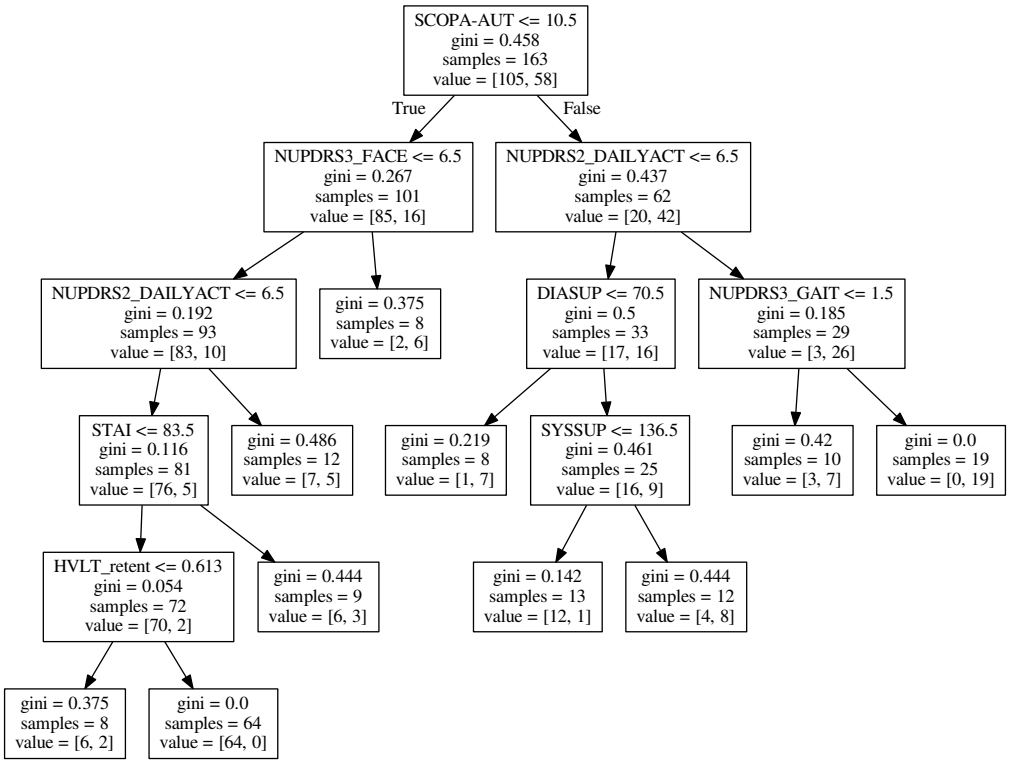

Figure A-7: **Decision tree** model of hybrid outcome in the 3-year trial setting (fold 1). See Figure A-6 for description.

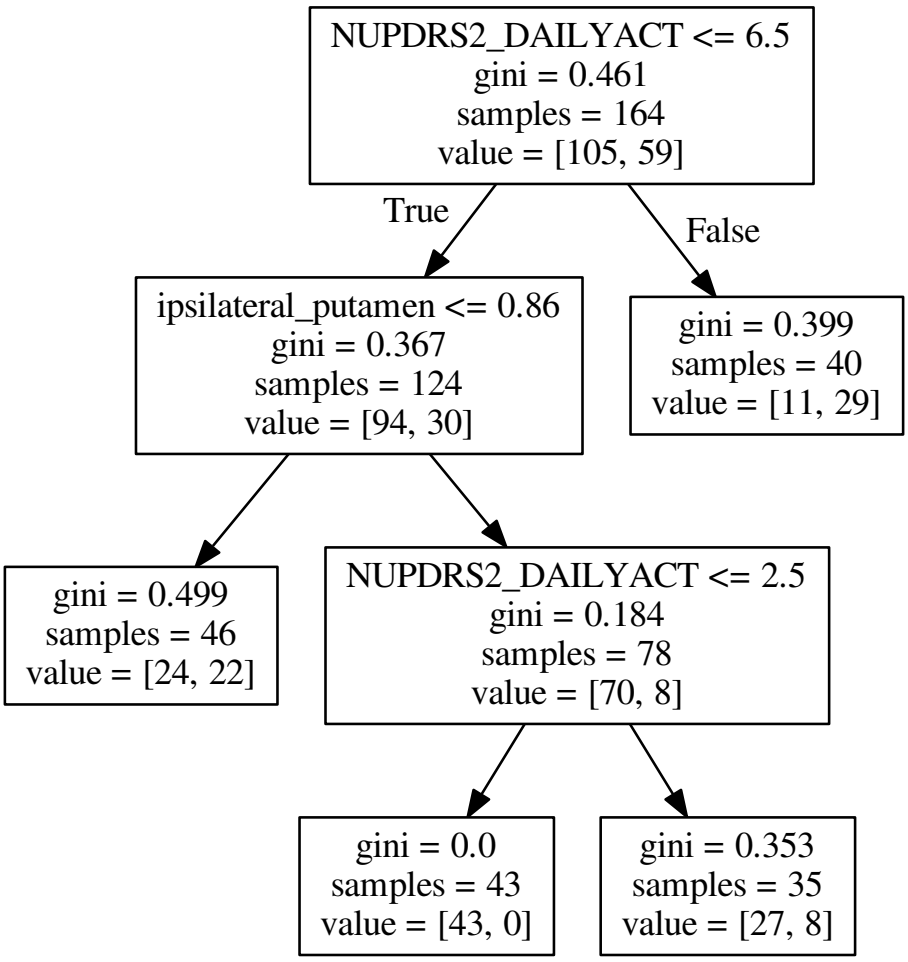

Figure A-8: **Decision tree** model of hybrid outcome in the 3-year trial setting (fold 2). See Figure A-6 for description.

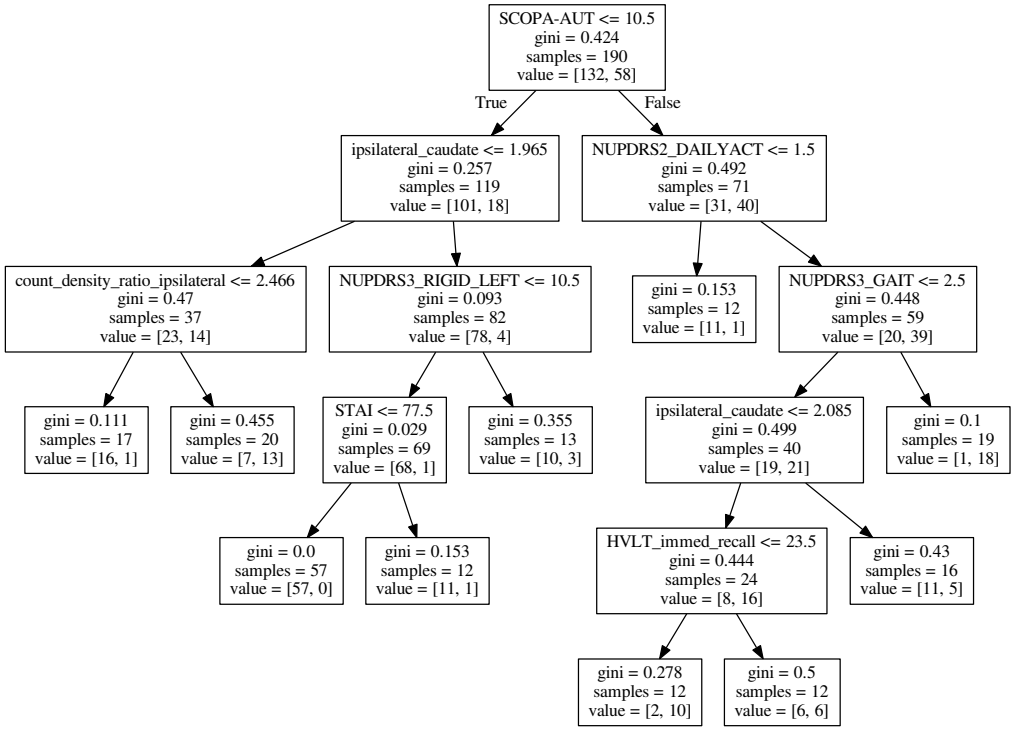

Figure A-9: **Decision tree** model of hybrid outcome in the 3-year trial setting (fold 3). See Figure A-9 for description.

Bibliography

[1] Simon Anders and Wolfgang Huber. Differential expression analysis for sequence count data. *Genome biology*, 11(10):R106, 2010.

[2] Inci M Baytas, Cao Xiao, Xi Zhang, Fei Wang, Anil K Jain, and Jiayu Zhou. Patient subtyping via time-aware lstm networks. In *Proceedings of the 23rd ACM SIGKDD international conference on knowledge discovery and data mining*, pages 65–74. ACM, 2017.

[3] Arthur L Benton, KD Hamsher, Nils R Varney, and O Spreen. *Judgment of line orientation*. Oxford University Press New York, 1983.

[4] Kevin M Biglan, David Oakes, Anthony E Lang, Robert A Hauser, Karen Hodgeman, Brittany Greco, Jillian Lowell, Rebecca Rockhill, Ira Shoulson, Charles Venuto, et al. A novel design of a phase iii trial of isradipine in early parkinson disease (steady-pd iii). *Annals of clinical and translational neurology*, 4(6):360–368, 2017.

[5] Jason Brandt. The hopkins verbal learning test: Development of a new memory test with six equivalent forms. *The Clinical Neuropsychologist*, 5(2):125–142, 1991.

[6] Chao Che, Cao Xiao, Jian Liang, Bo Jin, Jiayu Zho, and Fei Wang. An rnn architecture with dynamic temporal matching for personalized predictions of parkinson's disease. In *Proceedings of the 2017 SIAM International Conference on Data Mining*, pages 198–206. SIAM, 2017.

[7] Christopher S. Coffey, Chelsea J. Caspell-Garcia, and Eric D. Foster. Variable definitions and score calculations. Technical report, University of Iowa, Department of Biostatistics Clinical Trials Statistical and Data Management Center, 2017.

[8] Sarah F Cook and Robert R Bies. Disease progression modeling: key concepts and recent developments. *Current pharmacology reports*, 2(5):221–230, 2016.

[9] Cameron Davidson-Pilon, Jonas Kalderstam, Paul Zivich, Ben Kuhn, Andrew Fiore-Gartland, Luis Moneda, Gabriel, Daniel WIlson, Alex Parij, Kyle Stark, Steven Anton, Lilian Besson, Jona, Harsh Gadgil, Dave Golland, Sean Hussey, Ravin Kumar, Javad Noorbakhsh, Andreas Klintberg, Jakub Kaluzka, Isaac

Slavitt, Eric Martin, Eduardo Ochoa, Dylan Albrecht, dhuynh, Denis Zgonjanin, Daniel Chen, Chris Fournier, Arturo, and André F. Rendeiro. Camdavidson-pilon/lifelines: v0.21.2, May 2019.

[10] Eduardo De Pablo-Fernández, Andrew J Lees, Janice L Holton, and Thomas T Warner. Prognosis and neuropathologic correlation of clinical subtypes of parkinson disease. *JAMA neurology*, 76(4):470–479, 2019.

[11] Carl Doersch. Tutorial on variational autoencoders. *arXiv preprint arXiv:1606.05908*, 2016.

[12] Nir Dotan, Danit Mechlovich, Jennifer Yarden, and Martin Rabey. Validation of a blood based gene expression assay for monitoring disease severity and aid in the early diagnosis of parkinson's disease. Technical report, BioShai Ltd., 2017.

[13] Richard L Doty, Paul Shaman, and Michael Dann. Development of the university of pennsylvania smell identification test: a standardized microencapsulated test of olfactory function. *Physiology & behavior*, 32(3):489–502, 1984.

[14] Bradley Efron. The efficiency of cox's likelihood function for censored data. *Journal of the American statistical Association*, 72(359):557–565, 1977.

[15] Terry Ellis, James T Cavanaugh, Gammon M Earhart, Matthew P Ford, Kenneth B Foreman, and Leland E Dibble. Which measures of physical function and motor impairment best predict quality of life in parkinson's disease? *Parkinsonism & related disorders*, 17(9):693–697, 2011.

[16] Faraz Faghri, Sayed Hadi Hashemi, Hampton Leonard, Sonja W Scholz, Roy H Campbell, Mike A Nalls, and Andrew B Singleton. Predicting onset, progression, and clinical subtypes of parkinson disease using machine learning. *bioRxiv*, page 338913, 2018.

[17] Seyed-Mohammad Fereshtehnejad and Ronald B Postuma. Subtypes of parkinson's disease: what do they tell us about disease progression? *Current neurology and neuroscience reports*, 17(4):34, 2017.

[18] Simone Fleige, Vanessa Walf, Silvia Huch, Christian Prgomet, Julia Sehm, and Michael W Pfaffl. Comparison of relative mrna quantification models and the impact of rna integrity in quantitative real-time rt-pcr. *Biotechnology letters*, 28(19):1601–1613, 2006.

[19] Kristian Varden Gjerde, Bernd Müller, Geir Olve Skeie, Jörg Assmus, Guido Alves, and Ole-Bjørn Tysnes. Hyposmia in a simple smell test is associated with accelerated cognitive decline in early parkinson's disease. *Acta Neurologica Scandinavica*, 138(6):508–514, 2018.

[20] Christopher G Goetz, Werner Poewe, Olivier Rascol, Cristina Sampaio, Glenn T Stebbins, Carl Counsell, Nir Giladi, Robert G Holloway, Charity G Moore, Gregor K Wenning, et al. Movement disorder society task force report on the hoehn

and yahr staging scale: status and recommendations the movement disorder society task force on rating scales for parkinson's disease. *Movement disorders*, 19(9):1020–1028, 2004.

[21] Parkinson Study Group. Levodopa and the progression of parkinson's disease. *New England Journal of Medicine*, 351(24):2498–2508, 2004.

[22] Parkinson Study Group et al. Pramipexole vs levodopa as initial treatment for parkinson disease: a randomized controlled trial. *Jama*, 284(15):1931–1938, 2000.

[23] GE Healthcare. Datscan (ioflupane i 123 injection) for intravenous use, 2015.

[24] Miguel A Hernán. How to estimate the effect of treatment duration on survival outcomes using observational data. *bmj*, 360:k182, 2018.

[25] Dena G. Hernandez. Immunochip genotyping on dna samples from ppmi. Technical report, National Institute on Aging, Laboratory of Neurogenetics, 2017.

[26] Dena G. Hernandez. Selected genetic variants genotyped using neurox array. Technical report, National Institute on Aging, Laboratory of Neurogenetics, 2017.

[27] John R Hodges, Karalyn Patterson, Susan Oxbury, and Elaine Funnell. Semantic dementia: Progressive fluent aphasia with temporal lobe atrophy. *Brain*, 115(6):1783–1806, 1992.

[28] Margaret M Hoehn and Melvin D Yahr. Parkinsonism: onset, progression, and mortality. *Neurology*, 17(5):427–427, 1967.

[29] David Houghton, Howard Hurtig, Sharon Metz, Monique Giroux, Giselle Petzinger, Beth Fisher, Lauren Hawthorne, and Michael Jakowec. Parkinson's disease medications. 2017.

[30] TA Hughes, HF Ross, S Musa, S Bhattacherjee, RN Nathan, RHS Mindham, and EGS Spokes. A 10-year study of the incidence of and factors predicting dementia in parkinson's disease. *Neurology*, 54(8):1596–1603, 2000.

[31] Hirotaka Iwaki. Method to derive genotypes for loci associated with the risk of parkinson's disease and the summarized score for the genetic burden from whole genome sequenced data. Technical report, Laboratory of Neurogenetics, National Institute on Aging, 2017.

[32] jbiggsets (Education Testing Services). Factor analyzer. `https://github.com/EducationalTestingService/factor_analyzer`, 2019.

[33] Crispin Jenkinson, RAY Fitzpatrick, VIV Peto, Richard Greenhall, and Nigel Hyman. The parkinson's disease questionnaire (pdq-39): development and validation of a parkinson's disease summary index score. *Age and ageing*, 26(5):353–357, 1997.

[34] Murray W Johns. A new method for measuring daytime sleepiness: the epworth sleepiness scale. *sleep*, 14(6):540–545, 1991.

[35] Eric Jones, Travis Oliphant, Pearu Peterson, et al. Scipy: Open source scientific tools for python. 2001.

[36] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

[37] Jeanne C Latourelle, Michael T Beste, Tiffany C Hadzi, Robert E Miller, Jacob N Oppenheim, Matthew P Valko, Diane M Wuest, Bruce W Church, Iya G Khalil, Boris Hayete, et al. Large-scale identification of clinical and genetic predictors of motor progression in patients with newly diagnosed parkinson's disease: a longitudinal cohort study and validation. *The Lancet Neurology*, 16(11):908–916, 2017.

[38] Michael Lawton, Yoav Ben-Shlomo, Margaret T May, Fahd Baig, Thomas R Barber, Johannes C Klein, Diane MA Swallow, Naveed Malek, Katherine A Grosset, Nin Bajaj, et al. Developing and validating parkinson's disease subtypes and their motor and cognitive progression. *J Neurol Neurosurg Psychiatry*, 89(12):1279–1287, 2018.

[39] Daniel D Lee and H Sebastian Seung. Algorithms for non-negative matrix factorization. In *Advances in neural information processing systems*, pages 556–562, 2001.

[40] Bin Liu, Ying Li, Zhaonan Sun, Soumya Ghosh, and Kenney Ng. Early prediction of diabetes complications from electronic health records: A multi-task survival analysis approach. In *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.

[41] Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(Nov):2579–2605, 2008.

[42] Todd A MacKenzie, Jeremiah R Brown, Donald S Likosky, YingXing Wu, and Gary L Grunkemeier. Review of case-mix corrected survival curves. *The Annals of thoracic surgery*, 93(5):1416–1425, 2012.

[43] Angus D Macleod, Ingvild Dalen, Ole-Bjørn Tysnes, Jan Petter Larsen, and Carl E Counsell. Development and validation of prognostic survival models in newly diagnosed parkinson's disease. *Movement Disorders*, 33(1):108–116, 2018.

[44] Philipp Mahlknecht, Klaus Seppi, and Werner Poewe. The concept of prodromal parkinson's disease. *Journal of Parkinson's disease*, 5(4):681–697, 2015.

[45] Kenneth Marek. The parkinson's progression markers initiative. Technical report, Michael J. Fox Foundation, 2017.

[46] Kenneth Marek, Danna Jennings, Shirley Lasch, Andrew Siderowf, Caroline Tanner, Tanya Simuni, Chris Coffey, Karl Kieburtz, Emily Flagg, Sohini Chowdhury, et al. The parkinson progression marker initiative (ppmi). *Progress in neurobiology*, 95(4):629–635, 2011.

[47] Connie Marras and Anthony Lang. Parkinson's disease subtypes: lost in translation? *J Neurol Neurosurg Psychiatry*, 84(4):409–415, 2013.

[48] Connie Marras and Antony E Lang. Outcome measures for clinical trials in parkinson's disease: achievements and shortcomings. *Expert review of neurotherapeutics*, 4(6):985–993, 2004.

[49] Pablo Martínez-Martín, Carmen Rodríguez-Blázquez, Mario Alvarez, Tomoko Arakaki, Víctor Campos Arillo, Pedro Chaná, William Fernández, Nélida Garretto, Juan Carlos Martínez-Castrillo, Mayela Rodríguez-Violante, et al. Parkinson's disease severity levels and mds-unified parkinson's disease rating scale. *Parkinsonism & related disorders*, 21(1):50–54, 2015.

[50] David McGhee, Alexander Parker, Shona Fielding, John Zajicek, and Carl Counsell. Using 'dead or dependent'as an outcome measure in clinical trials in parkinson's disease. *J Neurol Neurosurg Psychiatry*, 86(2):180–185, 2015.

[51] Wes McKinney et al. Data structures for statistical computing in python. In *Proceedings of the 9th Python in Science Conference*, volume 445, pages 51–56. Austin, TX, 2010.

[52] Joyce A Mitchell, Jane Fun, and Alexa T McCray. Design of genetics home reference: a new nlm consumer health resource. *Journal of the American Medical Informatics Association*, 11(6):439–447, 2004.

[53] Brit Mollenhauer, Johannes Zimmermann, Friederike Sixel-Döring, Niels K Focke, Tamara Wicke, Jens Ebentheuer, Martina Schaumburg, Elisabeth Lang, Tim Friede, Claudia Trenkwalder, et al. Baseline predictors for progression 4 years after parkinson's disease diagnosis in the de novo parkinson cohort (denopa). *Movement Disorders*, 34(1):67–77, 2019.

[54] Mike A Nalls, Cornelis Blauwendraat, Costanza L Vallerga, Karl Heilbron, Sara Bandres-Ciga, Diana Chang, Manuela Tan, Demis A Kia, Alastair J Noyce, Angli Xue, et al. Expanding parkinson's disease genetics: novel risk loci, genomic context, causal insights and heritable risk. *BioRxiv*, page 388165, 2019.

[55] Ziad S Nasreddine, Natalie A Phillips, Valérie Bédirian, Simon Charbonneau, Victor Whitehead, Isabelle Collin, Jeffrey L Cummings, and Howard Chertkow. The montreal cognitive assessment, moca: a brief screening tool for mild cognitive impairment. *Journal of the American Geriatrics Society*, 53(4):695–699, 2005.

[56] Marlies Noordzij, Giovanni Tripepi, Friedo W Dekker, Carmine Zoccali, Michael W Tanck, and Kitty J Jager. Sample size calculations: basic principles and common pitfalls. *Nephrology dialysis transplantation*, 25(5):1388–1393, 2010.

[57] Alastair John Noyce, Andrew John Lees, and Anette-Eleonore Schrag. The prediagnostic phase of parkinson's disease. *J Neurol Neurosurg Psychiatry*, 87(8):871–878, 2016.

[58] Movement Disorder Society Task Force on Rating Scales for Parkinson's Disease. The unified parkinson's disease rating scale (updrs): status and recommendations. *Movement Disorders*, 18(7):738–750, 2003.

[59] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

[60] Emma Pierson, Pang Wei Koh, Tatsunori Hashimoto, Daphne Koller, Jure Leskovec, Nicholas Eriksson, and Percy Liang. Inferring multi-dimensional rates of aging from cross-sectional data. *arXiv preprint arXiv:1807.04709*, 2018.

[61] Werner Poewe, Klaus Seppi, Caroline M Tanner, Glenda M Halliday, Patrik Brundin, Jens Volkmann, Anette-Eleonore Schrag, and Anthony E Lang. Parkinson disease. *Nature reviews Disease primers*, 3:17013, 2017.

[62] Teun M Post, Jan I Freijer, Joost DeJongh, and Meindert Danhof. Disease system analysis: basic disease progression models in degenerative disease. *Pharmaceutical research*, 22(7):1038–1049, 2005.

[63] Ronald B Postuma and Daniela Berg. Advances in markers of prodromal parkinson disease. *Nature Reviews Neurology*, 12(11):622, 2016.

[64] Judith A. Potashkin and Jose A. Santiago. Whole blood rna biomarkers of parkinson's disease. Technical report, Rosalind Franklin University of Medicine and Science, The Cellular and Molecular Pharmacology Department, The Chicago Medical School, 2017.

[65] Judith A Potashkin, Jose A Santiago, Bernard M Ravina, Arthur Watts, and Alexey A Leontovich. Biosignatures for parkinson's disease and atypical parkinsonian disorders patients. *PloS one*, 7(8):e43595, 2012.

[66] Hude Quan, Bing Li, Chantal M Couris, Kiyohide Fushimi, Patrick Graham, Phil Hider, Jean-Marie Januel, and Vijaya Sundararajan. Updating and validating the charlson comorbidity index and score for risk adjustment in hospital discharge abstracts using data from 6 countries. *American journal of epidemiology*, 173(6):676–682, 2011.

[67] Jason DM Rennie and Nathan Srebro. Loss functions for preference levels: Regression with discrete ordered labels. In *Proceedings of the IJCAI multidisciplinary workshop on advances in preference handling*, pages 180–186. Kluwer Norwell, MA, 2005.

[68] Liana S Rosenthal, Daniel Drake, Roy N Alcalay, Debra Babcock, F DuBois Bowman, Alice Chen-Plotkin, Ted M Dawson, Richard B Dewey Jr, Dwight C German, Xuemei Huang, et al. The ninds parkinson's disease biomarkers program. *Movement Disorders*, 31(6):915–923, 2016.

[69] Clemens R Scherzer, Aron C Eklund, Lee J Morse, Zhixiang Liao, Joseph J Locascio, Daniel Fefer, Michael A Schwarzschild, Michael G Schlossmacher, Michael A Hauser, Jeffery M Vance, et al. Molecular markers of early parkinson's disease based on gene expression in blood. *Proceedings of the National Academy of Sciences*, 104(3):955–960, 2007.

[70] Anette Schrag, Marjan Jahanshahi, and Niall Quinn. What contributes to quality of life in patients with parkinson's disease? *Journal of Neurology, Neurosurgery & Psychiatry*, 69(3):308–312, 2000.

[71] Peter Schulam and Suchi Saria. A framework for individualizing predictions of disease trajectories by exploiting multi-resolution structure. In *Advances in Neural Information Processing Systems*, pages 748–756, 2015.

[72] Peter Schulam and Suchi Saria. Integrative analysis using coupled latent variable models for individualizing prognoses. *The Journal of Machine Learning Research*, 17(1):8244–8278, 2016.

[73] Javaid I Sheikh and Jerome A Yesavage. Geriatric depression scale (gds): recent evidence and development of a shorter version. *Clinical Gerontologist: The Journal of Aging and Mental Health*, 1986.

[74] Stephen T Sherry, M-H Ward, M Kholodov, J Baker, Lon Phan, Elizabeth M Smigielski, and Karl Sirotkin. dbsnp: the ncbi database of genetic variation. *Nucleic acids research*, 29(1):308–311, 2001.

[75] Aaron Smith. *Symbol digit modalities test*. Western Psychological Services Los Angeles, CA, 1982.

[76] Hossein Soleimani, Adarsh Subbaswamy, and Suchi Saria. Treatment-response models for counterfactual reasoning with continuous-time, continuous-valued interventions. *arXiv preprint arXiv:1704.02038*, 2017.

[77] Charles D Spielberger, Fernando Gonzalez-Reigosa, Angel Martinez-Urrutia, Luiz FS Natalicio, and Diana S Natalicio. The state-trait anxiety inventory. *Revista Interamericana de Psicologia/Interamerican Journal of Psychology*, 5(3 & 4), 2017.

[78] Glenn T Stebbins, Christopher G Goetz, David J Burn, Joseph Jankovic, Tien K Khoo, and Barbara C Tilley. How to identify tremor dominant and postural instability/gait difficulty groups with the movement disorder society unified parkinson's disease rating scale: comparison with the unified parkinson's disease rating scale. *Movement Disorders*, 28(5):668–670, 2013.

[79] Karin Stiasny-Kolster, Geert Mayer, Sylvia Schäfer, Jens Carsten Möller, Monika Heinzel-Gutenbrunner, and Wolfgang H Oertel. The rem sleep behavior disorder screening questionnaire—a new diagnostic instrument. *Movement disorders*, 22(16):2386–2393, 2007.

[80] PPMI Study Team and the Clinical Trials Coordination Center. Symptomatic therapy (st) visits. Technical report, University of Rochester, NY, 2017.

[81] Bruce Thompson. Factor analysis. *The Blackwell Encyclopedia of Sociology*, 2007.

[82] Tom N Tombaugh and Nancy J McIntyre. The mini-mental state examination: a comprehensive review. *Journal of the American Geriatrics Society*, 40(9):922–935, 1992.

[83] Stefan Van Der Walt, S Chris Colbert, and Gael Varoquaux. The numpy array: a structure for efficient numerical computation. *Computing in Science & Engineering*, 13(2):22, 2011.

[84] Stephanie M van Rooden, Willem J Heiser, Joost N Kok, Dagmar Verbaan, Jacobus J van Hilten, and Johan Marinus. The identification of parkinson's disease subtypes using cluster analysis: a systematic review. *Movement disorders*, 25(8):969–978, 2010.

[85] Charles S Venuto, Nicholas B Potter, E Ray Dorsey, and Karl Kieburtz. A review of disease progression models of parkinson's disease and applications in clinical trials. *Movement Disorders*, 31(7):947–956, 2016.

[86] Martine Visser, Johan Marinus, Anne M Stiggelbout, and Jacobus J Van Hilten. Assessment of autonomic dysfunction in parkinson's disease: the scopa-aut. *Movement disorders: official journal of the Movement Disorder Society*, 19(11):1306–1312, 2004.

[87] Xiang Wang, David Sontag, and Fei Wang. Unsupervised learning of disease progression models. In *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 85–94. ACM, 2014.

[88] Richard A Washburn, Kevin W Smith, Alan M Jette, and Carol A Janney. The physical activity scale for the elderly (pase): development and evaluation. *Journal of clinical epidemiology*, 46(2):153–162, 1993.

[89] David Wechsler. Wechsler adult intelligence scale–fourth edition (wais–iv). *San Antonio, TX: NCS Pearson*, 22:498, 2008.

[90] Daniel Weintraub, Staci Hoops, Judy A Shea, Kelly E Lyons, Rajesh Pahwa, Erika D Driver-Dunckley, Charles H Adler, Marc N Potenza, Janis Miyasaki, Andrew D Siderowf, et al. Validation of the questionnaire for impulsive-compulsive disorders in parkinson's disease. *Movement disorders: official journal of the Movement Disorder Society*, 24(10):1461–1467, 2009.

[91] Ruiping Xia and Zhi-Hong Mao. Progression of motor symptoms in parkinson's disease. *Neuroscience bulletin*, 28(1):39–48, 2012.

[92] Yanbo Xu, Yanxun Xu, and Suchi Saria. A bayesian nonparametric approach for estimating individualized treatment-response curves. In *Machine Learning for Healthcare Conference*, pages 282–300, 2016.

[93] Xi Zhang, Jingyuan Chou, Jian Liang, Cao Xiao, Yize Zhao, Harini Sarva, Claire Henchcliffe, and Fei Wang. Data-driven subtyping of parkinson's disease using longitudinal clinical records: A cohort study. *Scientific reports*, 9(1):797, 2019.