



Tel-Aviv University

The Raymond and Beverly Sackler Faculty of Exact Sciences

The Blavatnik School of Computer Science and AI

Optimized reweighting of ordinal scales: methodology and application to Parkinson's Disease

Thesis submitted in partial fulfillment of graduate requirements for
the degree "Master of Sciences" in Tel-Aviv University

School of Computer Science and AI

By

Assaf Benesh

Under the supervision of

Prof. Ron Shamir

August 2025

Acknowledgments

I want to share my deep appreciation with everyone who supported me while I worked on this thesis and made it possible for me to finish it. I could not have done it without you.

First and foremost, I would like to thank my supervisor, Prof. Ron Shamir, who made all of this possible. Being extremely professional and setting the highest standards in everything we did, yet having infinite patience and understanding in your mentoring, is a rare combination that I truly admire and appreciate, and do not take for granted. Also, doing a master's thesis while holding a full-time industry job—especially in the exact sciences—is no trivial task. Statistics show that most such attempts end before the thesis is completed, so accepting a student like me required a real leap of faith. Nevertheless, Ron believed in my ability to finish and welcomed me into his lab, which I cannot appreciate enough. Throughout the project, he stayed deeply involved—reviewing every output, suggesting new directions, providing resources, and connecting me with the right people—so the final work would be as complete and polished as it could be.

Second, I would like to thank NeuraLight, and specifically Eddy Ben-Ami, for supporting my studies and letting me devote part of my work hours to academic research, seminars, and conferences. Your support made this possible, and I'm glad my approaches are already being used at NeuraLight and changing how we look at clinical scales. Also, your research mentorship served me well both at Neuralight and in the work on this thesis, from breaking down hard and vague research questions to healthy critical thinking. Additionally, I would like to share my deep appreciation for everyone at Neuralight who taught me so much about Parkinson's disease. This clinical knowledge, for example disease symptoms, treatment and medication options or disease states, gave me a real edge in the thesis, allowing me to understand the numbers I was looking at and take some meaningful steps that I otherwise could not have taken.

I would like to deeply thank Prof. Anat Mirelman and Prof. Roy Alcalay, for reviewing my work and providing valuable feedback and insights from the PD clinical perspective, as well as sharing the BeaT-PD dataset which allowed me to externally validate my results and strengthen my claims in this work.

I would also like to thank everyone in Ron Shamir's ACGT lab for sharing your wisdom, asking questions and challenging my ideas, helping me navigate the computing resources, and — most importantly — sharing your cookies and brownies that made my time at TAU much sweeter.

Data science isn't possible without data, so I would like to express my appreciation to the Michael J. Fox Foundation for Parkinson's Research and all its funding partners for creating and maintaining the PPMI database used in this study.

I also want to thank my parents, Avi and Tzvia, who raised me to be both ambitious and curious. You are amazing, and I am so lucky to have grown up in such a wonderful family. I owe you for everything I've achieved (and will achieve).

Finally, I would like to thank Noa, my life partner, who gave up a lot of our quality time so I could complete this research, and always listened to me blurbing about it — even when I bored you to death. I love you.

I would like to dedicate this thesis to the late Prof. Nir Giladi, who was both a partner and a mentor in this study and sadly passed away during the project's development.

Abstract

Parkinson’s disease (PD) is a highly heterogeneous condition with symptoms spanning motor and non-motor domains. Standard clinical scales, such as the Movement Disorder Society’s Unified Parkinson’s Disease Rating Scale (MDS-UPDRS), are commonly used in clinical trials where disease status is carefully monitored. They often rely on simply summing item values, assuming uniform item importance and potentially conflating true disease progression with medication effects.

Here we propose a novel data-driven approach to learn optimized weights for individual items in such scales — both rater-based and self-reported — so that total scores better reflect the underlying disease trajectory. By leveraging large-scale longitudinal data from the Parkinson’s Progression Markers Initiative (PPMI) database, our methods identified which items and value increments most strongly indicate PD progression, down-weighting less informative items and excluding redundant ones. Our results show that learned weights substantially improve the monotonic relationship between total scores and clinical progression. We validated our weights using both held-out PPMI data and an independent dataset (BeaT-PD), demonstrating their robustness and generalizability. Applying our results in clinical trials may assist in increasing power and in reducing the number of participants required to demonstrate intervention effect. Moreover, even a modest subset of purely self-reported items, when weighted appropriately, can track progression nearly as effectively as the full instrument. These findings open avenues for streamlined assessments and for reduced patient and clinician burden.

Contents

Acknowledgments	2
Abstract	4
Introduction	8
1 Background	9
1.1 Parkinson's Disease	9
1.1.1 Characteristics	9
1.1.2 Diagnosis	10
1.1.3 Monitoring	11
1.1.4 Treatment	13
1.2 Clinical scales	17
1.2.1 MDS-UPDRS	17
1.2.2 MoCA	21
1.2.3 The S&E ADL Scale	23
1.2.4 Disease Milestones Approach	25
1.3 Computational Methods	27
1.3.1 Linear Programming	27
1.3.2 Quadratic Programming	28
1.3.3 Integer Programming and Mixed-Integer Programming	28
1.3.4 Item Response Theory	29
2 Methods	31
2.1 Preprocessing	31
2.1.1 Data	31

2.1.2	Filtering	31
2.1.3	Encoding	32
2.2	Evaluation	33
2.2.1	External validation	33
2.2.2	Full index vs self-reported index	34
2.3	Formulations and weight optimization	35
2.3.1	Overview	35
2.3.2	Terminology	36
2.3.3	Formulations maximizing the weighted difference	37
2.3.4	Formulations maximizing consistency	40
3	Results	44
3.1	Comparing the performance of the different approaches	44
3.2	Reducing the number of items	46
3.3	Additional validation using PPMI	47
3.3.1	Initiation of symptomatic therapy	47
3.3.2	Time to milestone	50
3.4	External Validation	50
3.5	Hardness of the computational problem	52
3.6	Implementation details	55
3.7	Tool and code availability	55
4	Discussion	56
5	Supplementary	61
5.1	List of Abbreviations	61

5.2	PPMI Data	61
5.3	The weights learned by each approach	62
5.4	Comparison of the indexes to external scales	63
5.5	Correlation of the indexes with the time to first milestone	65
5.6	External validation using self-reported items	66

Introduction

This thesis proposes a new methodology to improve ordinal scales, by using optimization algorithms. Such scales are commonly used in clinical practice. Our work is motivated by the MDS-UPDRS scale for Parkinson’s disease (PD), and all method development and testing is shown for it.

The thesis is organized as follows:

Section 1 provides comprehensive background information relevant for this thesis. The first part contains an overview of PD, including its clinical characteristics, diagnosis, monitoring and treatment strategies. The second part discusses the existing clinical scales and instruments commonly used to track PD progression. The third part provides background on computational methods relevant to the thesis, and on alternative approaches for scale optimization.

Section 2 presents the methods developed and applied in this research. It describes the data preprocessing steps, introduce consistency, the optimization function that we used, and the external validations we conducted. Finally, it describes in detail all the approaches we took to optimize consistency both directly and heuristically.

Section 3 presents the results of this study. It compares the performance of the various optimization approaches in terms of consistency. This section further evaluates the potential for simplifying clinical scales by reducing the number of required items without sacrificing accuracy. It also contains validation of our outcomes against external clinical measures as well as correlations with other clinical scales.

Section 4 provides an in-depth discussion of the implications of the findings and suggests avenues for future research.

The Supplementary section (Section 5) contains additional supporting information and data.

1 Background

1.1 Parkinson’s Disease

1.1.1 Characteristics

Parkinson’s disease (PD - see abbreviations list in Supplementary 5.1) is the second most common neurodegenerative disorder globally, following Alzheimer’s disease, with an estimated 10 million individuals affected worldwide [1]. Projections indicate that by 2050, the number of people living with PD could exceed 25 million, primarily due to an aging global population [2].

PD has a long and notable history since its first description by Dr. James Parkinson in his 1817 essay, *An Essay on the Shaking Palsy* [3], and research into its nature continues today. PD is widely recognized by its most visible symptom — a characteristic “resting” tremor of the hands — but the full range of symptoms is broad and complex. Motor symptoms often begin asymmetrically, and typically include slowness of movement (bradykinesia), muscle rigidity, and problems with balance and posture. Beyond motor issues, many individuals with PD experience non-motor symptoms such as disturbances in sleep, mood changes like depression or anxiety, and subtle shifts in cognition. These wide-ranging symptoms reflect the complexity of the disease, which stems primarily from the loss of dopamine-producing neurons in a region of the brain called the substantia nigra [4], which is a part of the basal ganglia.

Age plays a major role in the development of PD [5]. While it can sometimes occur in younger adults, the risk of PD increases significantly after the age of 60. It is also more common in men than in women [6], although scientists are still examining why these differences exist [7, 8]. Interestingly, certain lifestyle factors have been found to influence the chances of developing PD; for instance, smoking has been surprisingly correlated with a lower risk of the disease (even after adjusting for age-related mortality, indicating the association is not solely due to smokers dying before typical PD onset) [9].

As to PD causes, most cases are idiopathic, i.e. there is no clear or confirmed reason why

they occur [10]. However, genetics does play a role for some patients. Researchers have discovered specific genes—such as LRRK2, SNCA, and GBA—that when mutated are linked to an increased likelihood of developing PD [11–13]. Nevertheless, these genetic causes account for only 10% to 15% of the cases [14], leaving many questions unanswered about how environmental and other factors might interact with one’s biology.

Adding to the complexity, scientists increasingly suspect that what we currently label as “Parkinson’s disease” may actually be a collection of related disorders, or subtypes, that fall under the same umbrella term [15]. Each subtype could involve slightly different pathways in the brain or even different risk factors [16]. Ongoing research is focused on understanding these variations in order to provide insights into how PD begins and progresses. By unraveling these details, researchers hope to create more targeted ways to address and eventually prevent this far-reaching disease.

1.1.2 Diagnosis

Diagnosing PD often begins with a careful clinical assessment, since there is no single test or biomarker that definitively confirms the disorder [17]. In its earliest stages, PD can be subtle. Many people initially notice only mild symptoms that may resemble other conditions—making a confident diagnosis difficult. Consequently, movement disorder specialists rely on a range of criteria and observations to increase diagnostic accuracy.

One major guideline is the Movement Disorder Society (MDS) Clinical Diagnostic Criteria [18]. Under these criteria, the presence of bradykinesia is a core requirement. Bradykinesia should be observed along with at least one additional cardinal motor symptom: rest tremor or muscular rigidity. Over time, problems with balance and posture can also emerge, but these typically appear in more advanced stages of PD. Physicians also examine a patient’s overall medical history (including family history), evaluate any changes in facial expression or speech, and look for supporting signs such as improvement when given dopaminergic therapy. Although a positive response to medication is not expected in every PD case, it can strengthen the clinical impression of PD.

Certain “red flags” can indicate an alternate diagnosis or a Parkinson’s-plus syndrome, rather

than classical PD. For example, if significant issues with memory or cognitive function appear very early, that might suggest a different disorder, such as dementia with Lewy bodies [19]. Likewise, if symptoms progress unusually fast or start highly symmetric, or if patients have severe autonomic dysfunction (like major blood pressure drops when standing, bladder problems, or severe constipation) at disease onset, clinicians may suspect conditions such as multiple system atrophy [20]. Identifying these red flags helps doctors avoid misdiagnosis and ensures that individuals receive the most appropriate care.

Another dimension of PD diagnosis involves the concept of a prodromal phase [21, 22]. This refers to a period that may begin several years before the classical movement symptoms become obvious. During this prodromal stage, subtle indicators may appear, such as an impaired sense of smell (hyposmia), mood changes (depression or anxiety), or a sleep disorder known as REM sleep behavior disorder (RBD), where a person physically acts out their dreams. Although not everyone with these early symptoms will go on to develop PD, they do increase the likelihood of a future diagnosis [23–25]. Recognizing these non-motor signals can be particularly helpful, as it allows for closer monitoring of individuals who may be at a higher risk [26].

To support a clinical diagnosis, physicians may order brain imaging. A dopamine transporter (DaT) scan, for instance, assesses the level of dopamine function in the basal ganglia—a region central to movement control and significantly affected by PD [27]. While a DaT scan showing reduced dopamine activity in the relevant brain regions can reinforce a suspected PD diagnosis, it is not, by itself, conclusive [28]. Ultimately, no single test can definitively confirm PD; the gold standard remains a thorough clinical examination combined with an ongoing evaluation of how symptoms progress. Regular follow-up with a neurologist or a movement disorder specialist allows for adjustments if new signs emerge, and it helps ensure that what appears to be PD truly aligns with the clinical picture over the long term.

1.1.3 Monitoring

Monitoring and evaluating PD progression is essential for guiding clinical decisions and comparing research findings. Over the years, a variety of rating scales and assessment tools

have been developed to capture the disease’s diverse symptoms and degrees of severity. These tools standardize how clinicians document progression and response to treatment, helping ensure consistency both within a clinical setting and across different studies.

Hoehn and Yahr Scale. One of the earliest and most widely known staging systems is the Hoehn and Yahr (H&Y) scale, introduced in 1967 [29]. The H&Y scale categorizes PD into five stages based on the distribution and severity of motor symptoms—ranging from mild, unilateral symptoms (Stage 1) through bilateral involvement, and ultimately significant postural instability or complete dependence (Stage 5). This straightforward approach has the advantage of simplicity and clear cut-offs but focuses primarily on motor features, overlooking many non-motor symptoms of PD.

Parkinson’s Disease Rating Scale and Columbia University Rating Scale. Following the H&Y scale, other tools like the Parkinson’s Disease Rating Scale (PDRS) [30] and the Columbia University Rating Scale (CURS) [31] emerged. They provide more detailed breakdown of motor symptoms (tremor, rigidity, bradykinesia, and postural stability) and generate cumulative scores. By separating different motor features, these scales offer more nuanced clinical insights than the simpler H&Y stages, yet they still largely neglect non-motor aspects such as mood disturbances or cognitive changes.

Unified Parkinson’s Disease Rating Scale. In the late 1980s, the Unified Parkinson’s Disease Rating Scale (UPDRS) was introduced [32, 33]. It attempted to provide a more holistic picture of PD by including multiple parts: mentation, behavior, and mood; activities of daily living; motor examination; and treatment-related complications. This comprehensive structure made the UPDRS a widely accepted tool for clinical trials and day-to-day practice. Nevertheless, it did not fully address every potential facet of PD, and questions arose regarding inter-rater variability, as the scale left room for subjective interpretation in how evaluations were performed [34].

Movement Disorder Society Unified Parkinson’s Disease Rating Scale. To refine and modernize the UPDRS, a task force organized by the MDS launched a revised version in 2007–2008, called the MDS-UPDRS [35]. Building on the older scale’s multi-part structure, the MDS-UPDRS includes additional items focused on non-motor symptoms such

as sleep disturbances, cognitive changes, and autonomic dysfunction. A key innovation of the MDS-UPDRS is its detailed instructions for performing each examination and scoring each item. These guidelines were introduced specifically to reduce inter-rater variability. By standardizing how each assessment should be conducted, the MDS-UPDRS aims to improve consistency within clinics and between research centers, thus making it more reliable for tracking progression over time or comparing outcomes across studies.

In addition to these core rating scales, specialized tests help quantify specific PD-related issues. For instance, various measures exist for sleep disruptions, including assessments for RBD [36, 37]. Smell tests, such as the University of Pennsylvania Smell Identification Test (UPSIT), are commonly used to detect olfactory deficits that often appear well before classic motor symptoms [38]. To evaluate activities of daily living, the Schwab and England (S&E) scale measures the degree of independence a person retains in tasks like dressing and eating [39].

Cognitive function is another critical domain in PD, as many individuals experience mild cognitive impairment or more pronounced difficulties over time. The **Montreal Cognitive Assessment (MoCA)** is a commonly used tool to screen for mild cognitive impairment [40]. It tends to be more sensitive in capturing the subtle cognitive changes seen in PD than some older tests [41]. By applying these specialized scales alongside broader assessments (such as the MDS-UPDRS), clinicians gain a detailed perspective on both the motor and non-motor facets of PD.

1.1.4 Treatment

Treating PD is a complex and evolving process that aims to relieve symptoms, maintain quality of life, and preserve function for as long as possible. No single strategy works for every individual, and therapy often needs to be adjusted over time. Clinicians consider many factors—including age, symptoms, comorbidities, and personal preferences—when designing a treatment plan [42]. Although there are different approaches to the timing and sequence of medications, levodopa remains the cornerstone therapy for most patients with PD, especially once motor symptoms begin to significantly affect daily life.

PD progresses steadily (monotonically) over the course of the illness, and there are currently no proven disease-modifying treatments capable of slowing or stopping the underlying neurodegeneration [43]. As a result, all available therapies alleviate symptoms but do not alter the fundamental disease trajectory.

Levodopa and Its Role. Levodopa, often combined with carbidopa (to reduce peripheral side effects), is the most effective treatment for the motor symptoms of PD [44]. Levodopa crosses the blood-brain barrier and is converted into dopamine in the brain, helping to replenish the diminished dopamine levels that underlie many of the motor symptoms of PD. When a person first starts levodopa, lower doses can often achieve good symptom control. However, as PD progresses and more dopaminergic neurons are lost, higher or more frequent doses may be needed to maintain the same level of benefit. Over time, the therapeutic window (the dose range that effectively relieves symptoms without causing side effects) may narrow, posing new challenges in balancing symptom control with adverse effects [45].

One innovation that aims to stabilize levodopa’s delivery is extended-release formulations, which release the drug gradually over several hours [46]. These formulations can help smooth out fluctuations in blood levels of the medication, potentially reducing “wearing off” periods—times when symptoms return as the previous dose tapers off. Even so, many patients still need to take multiple doses of levodopa throughout the day. Other variations, such as orally disintegrating tablets and gel infusions (e.g., levodopa-carbidopa intestinal gel) [47, 48], offer different ways to manage dosage and absorption issues, especially in more advanced PD.

Adjunct Medications. While levodopa is central, additional medications can enhance or extend its benefits:

- Dopamine Agonists (e.g., pramipexole [49], ropinirole [50], rotigotine [51]) directly stimulate dopamine receptors, offering another route for managing PD symptoms [52]. These are sometimes used early in the disease to delay the introduction of levodopa, especially in younger patients. Over the long term, they can be added to levodopa therapy for more comprehensive control of symptoms.
- Monoamine Oxidase B (MAO-B) Inhibitors (e.g., selegiline [53], rasagiline [54], safi-

namide [55]) slow down the breakdown of dopamine within the brain, thereby prolonging the effect of both natural and medication induced dopamine [56]. They can be used as monotherapy in the early stages of PD or added to levodopa to enhance its effect.

- Catechol-O-methyl Transferase (COMT) Inhibitors (e.g., entacapone [57], opicapone [58]) block an enzyme that degrades levodopa, increasing the amount of levodopa available to the brain [59]. These are commonly prescribed alongside levodopa to stabilize its plasma levels and reduce “wearing off” intervals.
- Amantadine can help reduce dyskinesias (involuntary, writhing movements) that often develop in response to long-term levodopa use [60]. It also has some mild antiparkinsonian properties that can be helpful in early or later stages.

Although each class of PD medications can offer benefit, they also carry risks for side effects. For example, dopamine agonists may cause sleepiness, hallucinations, or impulse control disorders (such as compulsive gambling) [61], while MAO-B inhibitors can interact with other medications and cause insomnia or dizziness [62]. Balancing effectiveness and tolerability is a careful process that often takes trial, adjustment, and close collaboration between patient and clinician.

Fluctuations and “On/Off” Phenomenon. A major challenge in long-term levodopa therapy is the “on/off” phenomenon [63, 64]. During “on” periods symptoms are controlled, while “off” periods involve the return of PD symptoms as levodopa’s effect wanes. Initially, “off” times tend to occur at predictable intervals—often just before the next dose [65]. As PD advances, these fluctuations can become more unpredictable, placing a burden on patients’ daily routines and overall quality of life. One way to address these fluctuations is by increasing the frequency of levodopa doses or adding adjunct medications to smooth out levels of dopamine in the brain. In some instances, changing from short-acting to extended-release formulations or adding rescue therapies (like inhaled levodopa) for sudden “off” episodes may also provide relief [66].

Despite these strategies, finding the “perfect” blend of medications and timing remains an

ongoing art. Clinicians must constantly balance dose adjustments to achieve maximum symptom control while minimizing side effects such as dyskinesias, hallucinations, or orthostatic hypotension [67]. It can take several medication trials, dose modifications, and combination therapies to arrive at a regimen that feels optimal to the individual patient [68].

Treatment Onset and Timing. Not all clinicians and patients start levodopa therapy right away. Some prefer to delay dopaminergic treatment in milder or earlier stages—particularly in younger patients—due to concerns about dyskinesias or reduced medication efficacy over the long term [69]. In such cases, medications like dopamine agonists or MAO-B inhibitors might be used initially, providing moderate relief without introducing levodopa too soon. Conversely, many neurologists argue that there is little benefit in postponing levodopa if quality of life is compromised; the risk of unnecessarily enduring symptoms may outweigh concerns regarding future side effects [70]. Ultimately, the decision about when to start dopaminergic therapy is highly individualized and depends on a person’s clinical needs, symptom burden, and lifestyle considerations.

Deep Brain Stimulation. As PD progresses, managing motor fluctuations, dyskinesias, and medication side effects can become increasingly difficult. At this stage, advanced therapies like deep brain stimulation (DBS) can offer meaningful improvement for carefully selected patients. DBS involves implanting electrodes into specific brain regions (commonly the subthalamic nucleus or the globus pallidus interna) and connecting them to a device that delivers controlled electrical impulses [71]. This stimulation helps normalize abnormal brain signaling and can significantly reduce both “off” times and dyskinesias, allowing for lower doses of levodopa in many cases.

While DBS has revolutionized treatment for some individuals, it is not suitable for everyone. Proper patient selection—typically those with clear motor fluctuations, relatively intact cognitive function, and no major psychiatric complications—is critical to good outcomes.

Overall, treating PD is a dynamic process that requires continuous monitoring and adaptation of therapy. Through a combination of levodopa, adjunct medications, and potentially advanced interventions like DBS, many individuals living with PD can achieve meaningful relief of their symptoms. Nonetheless, the journey to find the right medication “cocktail”

and dose schedule can be lengthy and requires close cooperation between patients, caregivers, and clinicians.

1.2 Clinical scales

In the previous section we briefly discussed the clinical scales relevant to PD. Here we will elaborate more on MDS-UPDRS and MoCA, the scales that we optimized in this research. We will also discuss S&E ADL and disease milestones, which were used for evaluation of our results.

1.2.1 MDS-UPDRS

The MDS-UPDRS is one of the most comprehensive tools for evaluating the various signs, symptoms, and impacts of PD. Figure 1 shows the scales’s structure.

MDS-UPDRS consists of 65 items organized in four main parts:

1. Part I: Non-Motor Aspects of Experiences of Daily Living
 - Part IA assesses cognitive function, depression, apathy, and other self-perceived non-motor symptoms. This part is filled by the rater after questioning the patient or caregiver.
 - Part IB focuses on items such as pain, constipation and sleep problems, and is filled by the patient or caregiver themselves.
2. Part II: Motor Aspects of Experiences of Daily Living (self-reported). This section evaluates how motor symptoms—like tremor, slowness, or stiffness—affect daily activities such as eating, dressing, and handwriting. Patients or caregivers typically complete these items, with the clinician providing guidance as needed.
3. Part III: Motor Examination (rater-assessed). This is a clinician-rated physical examination of a patient’s motor function. Items can be divided into several domains:
 - Axial symptoms (e.g., speech, facial expression, posture, and gait)

MDS UPDRS Score Sheet

1.A	Source of information	<input type="checkbox"/> Patient <input type="checkbox"/> Caregiver <input type="checkbox"/> Patient + Caregiver	3.3b	Rigidity– RUE	
			3.3c	Rigidity– LUE	
Part I			3.3d	Rigidity– RLE	
1.1	Cognitive impairment		3.3e	Rigidity– LLE	
1.2	Hallucinations and psychosis		3.4a	Finger tapping– Right hand	
1.3	Depressed mood		3.4b	Finger tapping– Left hand	
1.4	Anxious mood		3.5a	Hand movements– Right hand	
1.5	Apathy		3.5b	Hand movements– Left hand	
1.6	Features of DDS		3.6a	Pronation- supination movements– Right hand	
1.6a	Who is filling out questionnaire	<input type="checkbox"/> Patient <input type="checkbox"/> Caregiver <input type="checkbox"/> Patient + Caregiver	3.6b	Pronation- supination movements– Left hand	
			3.7a	Toe tapping– Right foot	
1.7	Sleep problems		3.7b	Toe tapping– Left foot	
1.8	Daytime sleepiness		3.8a	Leg agility– Right leg	
1.9	Pain and other sensations		3.8b	Leg agility– Left leg	
1.10	Urinary problems		3.9	Arising from chair	
1.11	Constipation problems		3.10	Gait	
1.12	Light headedness on standing		3.11	Freezing of gait	
1.13	Fatigue		3.12	Postural stability	
Part II			3.13	Posture	
2.1	Speech		3.14	Global spontaneity of movement	
2.2	Saliva and drooling		3.15a	Postural tremor– Right hand	
2.3	Chewing and swallowing		3.15b	Postural tremor– Left hand	
2.4	Eating tasks		3.16a	Kinetic tremor– Right hand	
2.5	Dressing		3.16b	Kinetic tremor– Left hand	
2.6	Hygiene		3.17a	Rest tremor amplitude– RUE	
2.7	Handwriting		3.17b	Rest tremor amplitude– LUE	
2.8	Doing hobbies and other activities		3.17c	Rest tremor amplitude– RLE	
2.9	Turning in bed		3.17d	Rest tremor amplitude– LLE	
2.10	Tremor		3.17e	Rest tremor amplitude– Lip/jaw	
2.11	Getting out of bed		3.18	Constancy of rest tremor	
2.12	Walking and balance			Were dyskinesias present?	<input type="checkbox"/> No <input type="checkbox"/> Yes
2.13	Freezing			Did these movements interfere with ratings?	<input type="checkbox"/> No <input type="checkbox"/> Yes
3a	Is the patient on medication?	<input type="checkbox"/> No <input type="checkbox"/> Yes		Hoehn and Yahr Stage	
3b	Patient's clinical state	<input type="checkbox"/> Off <input type="checkbox"/> On	Part IV		
3c	Is the patient on levodopa?	<input type="checkbox"/> No <input type="checkbox"/> Yes	4.1	Time spent with dyskinesias	
3.C1	If yes, minutes since last dose:		4.2	Functional impact of dyskinesias	
Part III			4.3	Time spent in the OFF state	
3.1	Speech		4.4	Functional impact of fluctuations	
3.2	Facial expression		4.5	Complexity of motor fluctuations	
3.3a	Rigidity– Neck		4.6	Painful OFF-state dystonia	

Figure 1: The MDS-UPDRS scale [35]

- Rigidity (muscle stiffness in the patient’s limbs and neck)
 - Bradykinesia (slowness of movement, measured by tasks such as finger tapping or hand movements)
 - Tremor (both resting and action tremors are observed in different limb positions)
4. Part IV: Motor Complications (rater-assessed). This part targets complications from dopaminergic therapy, including motor fluctuations (such as “on” and “off” times) and dyskinesias. These items are usually based on patient reports and the clinician’s judgment regarding frequency, severity, and impact on quality of life.

Scoring System and Item Summation. Each item in the MDS-UPDRS is scored on a scale of 0 to 4, where:

- 0 typically indicates no symptoms or normal function
- 1 corresponds to minimal symptoms or slight impairment
- 2 represents mild impairment that is noticeable but may not strongly affect daily life
- 3 is moderate impairment
- 4 is severe impairment

Figure 2 shows an example of the scoring of one item. The final score is the sum of the items scores. It is important to note that this naive summation assumes linearity and equal increments. However, the criteria for different scores vary from one item to another in a nonlinear fashion. For example, in assessing facial expression, scoring “1” reflects reduced blinking while “2” represents partially masked face. Clinicians must carefully follow the detailed instructions to apply these criteria consistently, minimizing inter-rater variability.

The sum of item values is the overall total MDS-UPDRS score, which offers a snapshot of disease severity at a given point in time. However, in both research and clinical practice, certain sub-parts are sometimes emphasized over others, depending on the study’s goals or the clinical question. For instance, many clinical trials use the Part III motor score as a

3.14 GLOBAL SPONTANEITY OF MOVEMENT (BODY BRADYKINESIA)

Instructions to examiner: This global rating combines all observations on slowness, hesitancy, and small amplitude and poverty of movement in general, including a reduction of gesturing and of crossing the legs. This assessment is based on the examiner's global impression after observing for spontaneous gestures while sitting, and the nature of arising and walking.

0: Normal:	No problems.
1: Slight:	Slight global slowness and poverty of spontaneous movements.
2: Mild:	Mild global slowness and poverty of spontaneous movements.
3: Moderate:	Moderate global slowness and poverty of spontaneous movements.
4: Severe:	Severe global slowness and poverty of spontaneous movements.

Figure 2: An example of an MDS-UPDRS question in part III about bradykinesia and the detailed scoring instructions [35]

primary outcome measure to gauge improvements in motor function following an intervention [72–74].

Additional Recorded Information. Along with the four parts of the scale, the MDS-UPDRS collects additional clinical details that can be useful for interpreting scores or comparing patients. These include the patient's H&Y stage, whether the patient is taking levodopa (and if so, time since last dose) and whether the patient is currently in "on" or "off" state. Clinicians can use this contextual information to gain a better understanding of how the patient's current status fits into their broader PD journey and to interpret MDS-UPDRS scores more accurately over time.

Differences in Symptom Responsiveness. An important point about Part III is that different sub-items—axial, rigidity, bradykinesia, and tremor—may not all respond equally to dopaminergic treatments. For instance, tremor and bradykinesia often show a noticeable response to levodopa or dopamine agonists, whereas certain axial symptoms, such as gait instability or speech changes, can be more resistant [75].

Usage in Clinical Trials. The MDS-UPDRS total score—or, in many cases, the Part III motor subscore—has become one of the most frequently used primary endpoints in PD clinical trials. As of May 2025, over 300 registered interventional studies on ClinicalTrials.gov list MDS-UPDRS Part III as a primary outcome measure [76]. Researchers value it for its ability to capture a broad range of motor and non-motor domains. By focusing on a unified metric, studies can more consistently determine whether a new intervention provides meaningful benefit for individuals with PD.

The MDS-UPDRS has become a key tool in both research and practice. It provides an in-depth look at how PD affects individuals’ daily lives and tracks symptoms in a structured way. However, while better than previous scales, the MDS-UPDRS still lacks precision in detecting early stage PD progression [77], suffers from significant error [78], and only a big change in the score (at least 5 points) is considered clinically pertinent [79].

1.2.2 MoCA

The MoCA is a brief, standardized tool developed by Dr. Ziad Nasreddine to screen for mild cognitive impairment (MCI) and other cognitive deficits [40]. It has gained popularity in clinical and research settings for its ability to detect subtle changes in cognitive function. Figure 3 shows the instrument.

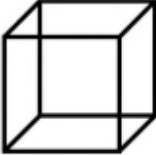
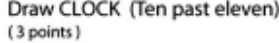
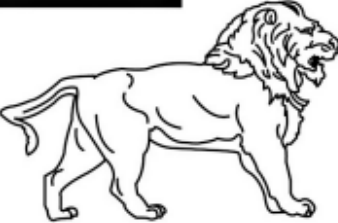
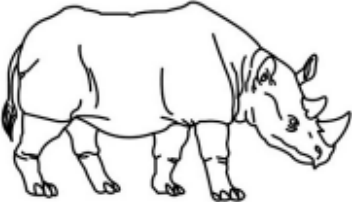
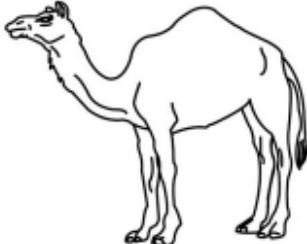
Test Structure The MoCA typically takes about 10–15 minutes to administer and covers several cognitive domains:

- **Visuospatial/Executive Function:** Assessed through tasks such as drawing a clock to a specific time or copying a geometric figure.
- **Naming:** Uses images (often animals) to evaluate the ability to identify and name objects.
- **Memory:** Involves immediate and delayed recall of a short list of words.
- **Attention:** Includes digit span, serial subtraction (e.g., 100 minus 7 repeatedly), and a vigilance task (tapping when a certain letter is heard).

MONTREAL COGNITIVE ASSESSMENT (MOCA)
Version 7.1 Original Version

NAME :
Education :
Sex :

Date of birth :
DATE :

VISUOSPATIAL / EXECUTIVE		 Copy cube  Draw CLOCK (Ten past eleven) (3 points)		POINTS																		
	<div style="text-align: center;">[]</div>	<div style="text-align: center;">[]</div>	<div style="display: flex; justify-content: space-around;"> <div style="text-align: center;">[] Contour</div> <div style="text-align: center;">[] Numbers</div> <div style="text-align: center;">[] Hands</div> </div>	___/5																		
NAMING		<div style="display: flex; justify-content: space-around; align-items: flex-end;"> <div style="text-align: center;">  [] </div> <div style="text-align: center;">  [] </div> <div style="text-align: center;">  [] </div> </div>			___/3																	
MEMORY	Read list of words, subject must repeat them. Do 2 trials, even if 1st trial is successful. Do a recall after 5 minutes.		<table border="1" style="width: 100%; border-collapse: collapse; text-align: center;"> <tr> <th></th> <th>FACE</th> <th>VELVET</th> <th>CHURCH</th> <th>DAISY</th> <th>RED</th> </tr> <tr> <td>1st trial</td> <td></td> <td></td> <td></td> <td></td> <td></td> </tr> <tr> <td>2nd trial</td> <td></td> <td></td> <td></td> <td></td> <td></td> </tr> </table>		FACE	VELVET	CHURCH	DAISY	RED	1st trial						2nd trial						No points
	FACE	VELVET	CHURCH	DAISY	RED																	
1st trial																						
2nd trial																						
ATTENTION	Read list of digits (1 digit/ sec.).		Subject has to repeat them in the forward order [] 2 1 8 5 4 Subject has to repeat them in the backward order [] 7 4 2		___/2																	
Read list of letters. The subject must tap with his hand at each letter A. No points if ≥ 2 errors		[] FBACMNAAJKLBAFAKDEAAAJAMOF AAB			___/1																	
Serial 7 subtraction starting at 100		[] 93 [] 86 [] 79 [] 72 [] 65 4 or 5 correct subtractions: 3 pts , 2 or 3 correct: 2 pts , 1 correct: 1 pt , 0 correct: 0 pt			___/3																	
LANGUAGE	Repeat : I only know that John is the one to help today. []		The cat always hid under the couch when dogs were in the room. []		___/2																	
Fluency / Name maximum number of words in one minute that begin with the letter F		[] _____ (N ≥ 11 words)			___/1																	
ABSTRACTION	Similarity between e.g. banana - orange = fruit		[] train - bicycle [] watch - ruler		___/2																	
DELAYED RECALL	Has to recall words WITH NO CUE	FACE []	VELVET []	CHURCH []	DAISY []	RED []	Points for UNCUED recall only	___/5														
Optional	Category cue																					
	Multiple choice cue																					
ORIENTATION	[] Date [] Month [] Year [] Day [] Place [] City						___/6															
© Z.Nasreddine MD		www.mocatest.org		Normal ≥ 26 / 30		TOTAL ___/30 Add 1 point if ≤ 12 yr edu																

Administered by: _____

Figure 3: The MoCA instrument [40]

- **Language:** Evaluates sentence repetition and verbal fluency (e.g., generating words that begin with a given letter).
- **Abstraction:** Assessed by asking how two words (such as "banana" and "orange") are alike.
- **Orientation:** Determines awareness of the current date, place, and situation.

The MoCA has a total maximum score of 30 points. Individuals with 12 years of education or less are often granted one extra point to their total score, aiming to compensate for disparities in formal education. A cutoff score of 26 or higher is generally deemed "normal", but this may vary depending on the population and purpose of the assessment.

Practice Effect and Localization A known limitation of the MoCA is the *practice effect*: if an individual repeats the test within a short time frame, previous familiarity with tasks can lead to an artificial improvement [80, 81]. Also, the test has been translated into multiple languages and culturally adapted, but these modifications usually require additional validation as cultural and linguistic nuances may influence comprehension [82–84]. Using validated local versions can improve accuracy and fairness in interpreting the results.

In summary, the MoCA is an efficient tool that samples several cognitive domains and can detect even mild cognitive impairment. Despite challenges such as practice effects and the need for proper localization, the MoCA remains a widely used instrument for early cognitive assessment.

1.2.3 The S&E ADL Scale

The S&E ADL scale is a simple yet widely used method to gauge how independently an individual with PD can manage everyday tasks. Developed by J.F. Schwab and A.C. England [39], it has been incorporated into many PD evaluations to complement other motor and non-motor assessments.

Scale Structure: The S&E scale is expressed as a percentage, ranging in 10% decrements from 100% (complete independence) down to 0% (bedridden or no activity possible). A rating of 100% typically indicates full function, with no difficulty in performing daily activities such as dressing, eating, or managing finances. Scores in the 70%–80% range reflect mild difficulties; the person can still manage most tasks but may require extra time or effort. When scores drop to around 50%, assistance for some tasks becomes necessary, and daily life is noticeably affected by symptoms. At 20%–30%, individuals generally need significant help for most activities and may no longer be able to live independently. A score of 0% signifies complete dependence, often with the individual confined to bed and unable to carry out even basic self-care.

Administration and Use. The S&E scale can be completed by either the clinician, the patient, or both in collaboration. Because it focuses on real-world task performance, it provides a rapid overview of how motor and non-motor symptoms manifest in daily life. In research, the S&E scale is often used alongside other scales (e.g., MDS-UPDRS) to correlate functional independence with more detailed motor or cognitive assessments.

Advantages and Limitations. The S&E ADL measure is quick to administer and requires minimal instructions, making it practical in both clinical and research settings. It provides a straightforward, intuitive percentage format that clearly conveys a patient's overall functional capacity, which holds value for both clinicians and patients. However, it has notable limitations. It does not offer fine-grained insights into which specific activities pose the greatest challenge. Additionally, the measure can be somewhat subjective, as ratings hinge on patient or rater perception of "difficulty" or independence. Scores may also be influenced by non-PD factors, such as comorbidities, the patient's living environment, or the level of available support.

Overall, the S&E ADL scale offers a valuable "snapshot" of functional independence. It complements more detailed rating scales and provides an easily communicated metric of how PD affects everyday life.

1.2.4 Disease Milestones Approach

A recent strategy for monitoring PD progression identified 25 specific "milestones" that mark meaningful changes in clinical status. As described by Brumm et al. [85], this milestone-based framework was devised within the PPMI cohort to capture distinct, functionally relevant events in several symptom domains. The underlying premise is that each milestone reflects a level of severity likely to indicate a notable shift in how PD affects daily life. Figure 4 shows these milestones' definition. The milestones domains are walking and balance, motor complications, cognition, autonomic dysfunction, functional dependence and activities of daily living. Each milestone is triggered by meeting strict criteria (e.g., certain MDS-UPDRS item scores). These cutoffs are intended to ensure that only clearly "clinically meaningful" events are captured, thereby reducing ambiguity about whether a given change truly reflects a significant progression.

Scoring and Practical Considerations Brumm et al. defined the **composite milestone endpoint** as meeting any one (or more) of the listed criteria. Because these milestones are chosen to represent disability rather than mild or easily reversible changes, this method aims to be less susceptible to temporary symptomatic improvement.

Advantages and Limitations. The milestone-based approach emphasizes clinically meaningful changes, reflecting the onset or worsening of disability rather than minor score fluctuations. By integrating both motor and non-motor domains, it offers a composite view of disease progression and remains relatively unaffected by medication state—particularly in areas like cognition and functional dependence—making it suitable for pragmatic trial designs. However, its complexity can hinder precise interpretation, especially as some milestones may recur or partially remit. While useful for identifying shifts into more advanced stages, this method may miss subtler, subclinical changes.

In summary, disease milestones provide an alternative perspective on PD staging and progression. Rather than relying solely on continuous scale scores, these milestones represent key points at which patients experience functionally significant changes. As reported in [85], the milestone approach appears promising for both observational studies and clinical trials aiming to evaluate interventions with a focus on meaningful, patient-relevant outcomes.

Criteria used to define progression milestones		
Progression milestone	Assessment(s)	Criteria
Domain 1: Walking and balance		
Walking and balance	MDS-UPDRS item 2.12	Response ≥ 3
Freezing	MDS-UPDRS item 2.13	Response ≥ 3
Gait	MDS-UPDRS item 3.10	Response ≥ 3 (ON or OFF)
Freezing of gait	MDS-UPDRS item 3.11	Response = 4 (ON or OFF)
Postural instability	MDS-UPDRS item 3.12	Response ≥ 3 (ON or OFF)
Hoehn and Yahr stage	Hoehn and Yahr Stage	Response ≥ 4 (ON or OFF)
Domain 2: Motor complications		
Dyskinesias	MDS-UPDRS items 4.1 and 4.2	Response ≥ 3 (on <i>both</i> items)
Fluctuations (functional impact)	MDS-UPDRS item 4.4	Response ≥ 3
Fluctuations (complexity)	MDS-UPDRS item 4.5	Response ≥ 3
Domain 3: Cognition		
Cognitive impairment (MoCA)*	MoCA	Score < 21
Cognitive impairment (MDS-UPDRS)	MDS-UPDRS item 1.1	Response ≥ 3
Hallucinations	MDS-UPDRS item 1.2	Response ≥ 3
Apathy	MDS-UPDRS item 1.5	Response ≥ 3
Dementia (clinical diagnosis)*	Site investigator assessment	PDD (per Investigator)
Dementia (composite)*	(1) Cognitive testing (2) Site investigator assessment	Impairment [†] on ≥ 2 cognitive domains; and Functional impairment (per investigator)
Domain 4: Autonomic dysfunction		
Urinary incontinence**	(1) MDS-UPDRS item 1.10 (2) SCOPA-AUT items 8 and 9	Response ≥ 3 ; and Response ≥ 2 (on <i>either</i> item)
Orthostatic hypotension**	(1) SCOPA-AUT item 15 (2) Systolic blood pressure (3) Diastolic blood pressure	Response ≥ 2 ; and Change of ≥ 20 mm Hg (sitting to standing); and Change of ≥ 10 mm Hg (sitting to standing)
Syncope (MDS-UPDRS)	MDS-UPDRS item 1.12	Response = 4
Syncope (SCOPA-AUT)**	SCOPA-AUT item 16	Response ≥ 1
Domain 5: Functional dependence		
Schwab & England	Schwab & England	Response < 80
Domain 6: Activities of daily living		
Choking	MDS-UPDRS item 2.3	Response ≥ 3
Eating	MDS-UPDRS item 2.4	Response ≥ 3
Dressing	MDS-UPDRS item 2.5	Response ≥ 3
Hygiene	MDS-UPDRS item 2.6	Response ≥ 3
Speech	MDS-UPDRS item 3.1	Response ≥ 3 (ON or OFF)

Figure 4: The definitions of milestones used in Brumm et al. [85]

1.3 Computational Methods

In this section we describe briefly some basic background on the formulations and algorithms used in this study.

1.3.1 Linear Programming

The linear programming (LP) problem aims to optimize a linear objective function subject to a set of linear constraints [86]. A typical LP can be written in the following canonical form:

$$\begin{aligned} & \text{maximize} && c^T x \\ & \text{subject to} && Ax \leq b \\ & && x \in \mathbb{R}^n \end{aligned} \tag{1}$$

where $x \in \mathbb{R}^n$ is the vector of decision variables, $c \in \mathbb{R}^n$ defines the coefficients of the objective function, $A \in \mathbb{R}^{m \times n}$ is a matrix of constraint coefficients, and $b \in \mathbb{R}^m$ is the corresponding right-hand side vector. In addition to inequalities of the form $Ax \leq b$, other types of linear constraints such as $Ax \geq b$ or $Ax = b$ can be accommodated by simple transformations.

Linear programs are exceptionally versatile, allowing a diverse array of real-world problems to be modeled effectively, including resource allocation, transportation, network flow, scheduling, and other combinatorial optimization subproblems. The essential requirement is that both the objective function and the constraints must be linear in terms of the decision variables.

Importantly, linear programs can be solved in polynomial time. Historically, the *simplex algorithm* [87] was the first widely adopted LP algorithm. Although the simplex method demonstrates worst-case exponential running time on certain constructed problem instances, it is often extremely efficient for practical problems. On the other hand, *interior-point methods*, such as Karmarkar’s algorithm [88], admit proven polynomial-time complexity and thus establish that LP lies within the complexity class **P**. Modern LP solvers typically combine both simplex-based approaches (which tend to exploit the structure of large-scale, sparse problems) and interior-point methods (which can be very effective for more dense or ill-conditioned instances).

1.3.2 Quadratic Programming

Quadratic programming (QP) extends the framework of LP by allowing a quadratic objective function, while the constraints typically remain linear [89]. A canonical form of a quadratic program can be written as:

$$\begin{aligned} & \text{minimize} && \frac{1}{2}x^T Q x + c^T x \\ & \text{subject to} && Ax \leq b \\ & && x \in \mathbb{R}^n \end{aligned} \tag{2}$$

where $x \in \mathbb{R}^n$ is the vector of decision variables, $Q \in \mathbb{R}^{n \times n}$ is a matrix that defines the quadratic portion of the objective function, $c \in \mathbb{R}^n$ represents the linear component of the objective, $A \in \mathbb{R}^{m \times n}$ is a matrix of constraint coefficients, and $b \in \mathbb{R}^m$ is the right-hand-side vector for the constraints. Additional constraints such as $Ax \geq b$ or $Ax = b$ can be introduced similarly to LP. The factor of $\frac{1}{2}$ in the objective is largely conventional, facilitating simpler gradient expressions in optimization algorithms.

A critical feature of QP is the definiteness of the matrix Q . If Q is *positive semidefinite* (PSD), the objective function is convex, and the problem is termed a *convex quadratic program* [90]. Convex QPs can be solved in polynomial time, but if Q is indefinite (contains both positive and negative eigenvalues), then the problem is generally *non-convex* and can be NP-hard [91].

Several heuristic approaches are commonly employed for solving QPs, differing in efficiency, numerical stability, and suitability for particular problem structures. These include active set methods [92], interior point methods [93], gradient based methods [94–96] and branch-and-bound [97].

1.3.3 Integer Programming and Mixed-Integer Programming

Integer programming (IP) extends LP by restricting some or all of the decision variables to be integers [98]. A general *integer linear program* (ILP) can be written in the following

form:

$$\begin{aligned}
& \text{minimize} && c^T x \\
& \text{subject to} && Ax \leq b \\
& && x \in \mathbb{Z}^n
\end{aligned} \tag{3}$$

where $x \in \mathbb{Z}^n$ indicates that each component of x must be an integer, $c \in \mathbb{R}^n$ is the cost vector, and $A \in \mathbb{R}^{m \times n}$, $b \in \mathbb{R}^m$ define the linear constraints. Notably, if only a subset of the variables need to be integral, the problem is called a *mixed-integer linear program* (MILP).

In such cases,

$$x = \begin{pmatrix} x_I \\ x_C \end{pmatrix}$$

with $x_I \in \mathbb{Z}^k$ (integer variables) and $x_C \in \mathbb{R}^{n-k}$ (continuous variables).

Solving a general ILP is *NP-hard* [99]. Hence, ILP and MILP solvers typically rely on exponential-time techniques in the worst case, though many specialized algorithms and heuristics exploit problem structure to achieve reasonable performance in practice [100].

Key algorithmic frameworks employed in ILP and MILP include branch-and-bound methods [101, 102], cutting plane methods [103], branch-and-cut [104] and various heuristics and metaheuristics [105–109].

1.3.4 Item Response Theory

Item Response Theory (IRT) is a family of statistical models and methods widely used in psychometrics. At its core, IRT aims to relate the probability of a respondent (e.g., an examinee) providing a particular response (e.g., a correct answer) to both the respondent’s latent trait level and certain characteristics of the test items themselves. In a typical IRT model, each respondent’s ability or latent trait is represented by a single numerical value $\theta \in \mathbb{R}$. Each item has its own parameters that describe how it discriminates among different ability levels and how challenging or difficult it is. One of the simplest and most frequently used IRT models is the *Rasch model* [110] (also known as the 1-parameter logistic model), defined as:

$$P(\text{correct} \mid \theta_i, b_j) = \frac{1}{1 + e^{-(\theta_i - b_j)}} \tag{4}$$

where θ_i is the ability level of person i and b_j is the difficulty parameter of item j .

Variants of this formulation allow for additional item parameters, such as *discrimination* and *guessing* (in 2-parameter and 3-parameter logistic models), leading to more flexible curves [111–113].

Estimation and Implementation. IRT models often involve estimating both person and item parameters from observed response data. This estimation is typically carried out via maximum likelihood or Bayesian methods, using algorithms such as the Expectation-Maximization (EM) approach [114] or Markov Chain Monte Carlo (MCMC) in a Bayesian context [115, 116].

2 Methods

In this section we will discuss our data and the various methods that we developed and applied.

2.1 Preprocessing

2.1.1 Data

We used data from the PPMI [117] - an international, multi-center longitudinal study aimed at identifying biomarkers of PD progression. While PPMI contains a variety of data types including imaging and genetic data, in this study we only used MDS-UPDRS [118] for summarizing patients' clinical state and MoCA [40] for their cognition, as well as a few other examination types for validation. The cohort contained $>12,000$ MDS-UPDRS tests of $\sim 1,500$ PD patients. See supplementary 5.2 for more details.

2.1.2 Filtering

The MDS-UPDRS [35] is a 65-item scale divided into four parts (Part I: non-motor experiences of daily living, II: motor experiences of daily living, III: clinician-rated motor examination, IV: motor complications. See Section 1.2.1). As our input, we used the 59 questions in parts I, II and III as well as MoCA. While PPMI contains various types of subjects (Healthy, PD, prodromal PD and other disorders) we focused only on PD patients in this analysis, and in particular excluded prodromal patients. We also removed examinations where the rater noted that dyskinesia interfered with the rating.

Since we wanted our tool to be applicable in regular clinical visits, we excluded visits where the PD patients were measured in 'OFF' state, as this kind of measurement often requires patients to purposely stop taking their medications and thus introduces undesired burden on them. See 'Discussion' section where we address this decision.

Finally, we removed the baseline visit of each patient from our analysis, as we suspect the first visit might be biased due to the Hawthorn effect [119], as the act of joining a clinical

trial by itself might create some temporal positive "improvement" in the patients state, compared to followup visits.

After filtering the data we had a total of 3,295 examinations for 711 distinct patients (averaging in 4.63 exams per patient, with median time difference of one year between consecutive visits). Note the data is not distributed evenly across PD severity levels, and is heavily biased towards early patients - which, as we shall argue later, is beneficial for the analyses.

Table 1 presents the characteristics of the final cohort after data cleaning.

Characteristic	Mean	Std	Min	Max
Age (years)	63.29	9.6	33.2	85.9
Disease duration (years)	2.54	1.9	0.92	14.09
MDS-UPDRS total score	36.65	16.62	4	122
H&Y Stage	1.78	0.55	0	4
MoCA total score	26.22	3.25	6	30
Gender (M%)	61.6%			
Follow-up time (years)	4.78	3.23	0.5	12.08
Number of visits	4.63	2.47	2	12

Table 1: PPMI participants characteristics. Results are shown for PPMI cohort after filtering (3,295 examinations for 711 PD patients). The first five characteristics listed correspond to each patient’s first visit included in the analysis.

2.1.3 Encoding

To make the data canonic and usable for the next step, we transformed it as follows. First, while MDS-UPDRS gives higher scores for more severe patients, the MoCA score decreases with severity from 30 to 0 - patients get full points for correct answers. To have both scales monotone increasing with severity, we flipped the values of each MoCA item such that the value is the number of points deducted instead of the number of points gained.

Next, for each question, we assigned a binary variable for each unit increment in the answer. For example, an MDS-UPDRS question that can have an answer between 0 and 4 was

transformed into four binary variables x_1, x_2, x_3, x_4 , where $x_i = 1$ if the answer is at least i . Hence, the answer 0 is mapped to $[0,0,0,0]$, 1 is mapped to $[1,0,0,0]$, 2 is mapped to $[1,1,0,0]$, 3 is mapped to $[1,1,1,0]$ and 4 is mapped to $[1,1,1,1]$. This way, for example, the answers to the 59 questions used from the MDS-UPDRS are represented by 236 binary variables. This type of encoding for ordinal data is sometimes referred to as thermometer encoding [120] or cumulative binary encoding. By giving non-negative weights to items, w_1, w_2, w_3, w_4 , the score of a question $\sum_i w_i \cdot x_i$ is monotone non-decreasing: Higher answers are assigned higher scores. The total weighted sum of all answers in a patient’s visit is called its *progression index*.

2.2 Evaluation

We split the data into 80% training set and 20% test/evaluation set, such that no patient appears in both train and test sets. The learning of weights was done only on the training set, and evaluated on the test set (note that splitting by site was not possible as recruitment site information is not available in PPMI).

Our primary metric for assessing the optimized weights was the percentage of visit pairs for the same patient in which the later visit received a higher progression index. We call this metric *consistency*. A score with higher consistency is better. We also measured the number of non-zero weights assigned to items. A lower number reflects a simpler scale that is easier to implement.

2.2.1 External validation

We compared the performance of the computed progression index against external progression criteria, and tested whether it performs better than the baseline approaches. The first set of criteria were based on data available in PPMI. First, we examined the relationship between a visit’s score and the time elapsed from that visit until the start of levodopa treatment, assuming an effective scale should assign higher scores to patients who are closer to beginning treatment. Second, we checked the scores concordance with the Schwab and

England Activities of Daily Living (S&E ADL) scale [39], expecting a negative correlation between our disease progression score and the ADL score. Lastly, we used the milestones defined by Brumm et al. [85] and checked how well our progression index predicts the time it would take a patient to reach the first milestone. We tested 20 out of the 25 milestones defined in [85], for which a sufficiently large fraction of the visits had data. We assumed a good index should exhibit a strong negative correlation, so that higher scores are associated with a shorter time to reaching the first milestone.

Finally, we validated the consistency of our weights against an additional, external cohort of PD patients obtained from the BeaT-PD project (204-16TLV) [121]. The BeaT-PD cohort included 300 recently diagnosed patients with PD (mean age 61.67 ± 10.34 years with mean disease duration of 2.5 ± 1.1 years) who were clinically and genetically assessed over 5 years. After applying filtering criteria similar to those used for the PPMI dataset, as described in Section 2.1.2 - but without removing baseline visits, to preserve dataset size - we retained 79 patients with a total of 201 visits. For the validation of the self-report index we applied a milder filtering approach, and did not exclude visits based on MDS-UPDRS part 3 criteria (clinical state or dyskinesia interference), as these are not self-reported measures.

2.2.2 Full index vs self-reported index

We also developed an index that uses only MDS-UPDRS questions that are self-reported and do not require a trained rater. This index uses only the patient’s questionnaire (the second half of part I and the entire part II). We tested if we can develop an index with good results that uses only self-reported items. Such data is significantly easier and cheaper to collect than a rater’s evaluation. It can be regularly collected in every visit of a patient to the clinic, and even via a remote application that the patient can operate from home.

2.3 Formulations and weight optimization

2.3.1 Overview

We developed a variety of formulations for optimizing the weights in the scale. The first set of approaches seek to maximize objective functions that are similar to — but not identical to — the consistency measure, are justified by a solid rationale, and can be optimized efficiently. Empirically, they can be solved to optimality on our data within a few minutes of computation on a standard laptop. These approaches include:

- **MeanDiff** - maximizing the mean difference between pairs of visits of the same patient, across all patients.
- **MeanDiff-W** (Weighted) - similar to the above, but penalizing more for negative differences, corresponding to pairs of visits for which the score decreased. The objective is to maximize the weighted sum of differences.
- **MeanDiff-QP** (Quadratic Penalty) - similar the former but introducing quadratic penalty for decreases - thus penalizing larger decreases more heavily. The objective is to maximize the sum of differences while minimizing the penalty.
- **MeanDiff-SV** (Small Variance) - similar to the MeanDiff approach, with an additional penalty factor measuring the variance of score differences between visits. The objective is to maximize the mean difference while minimizing the differences' variance, incentivizing stable increases.

For each of the approaches above, we also added an optional regularization term for minimizing the number of non-zero weights, incentivizing sparse solutions. This was both a goal by itself (as discussed earlier), and was also beneficial to prevent overfitting the training data.

Our second set of approaches aim to optimize consistency. They seek weights that will maximize the number of consistent pairs. We considered two variants of this problem: one where weights can have any real value, and one where only integer weights are allowed. We call these formulations **Cons** and **Cons-Int**, respectively. The objective functions in these

formulations are not convex, so finding the global optimum is computationally harder. We used algorithms that may take exponential time to reach an optimum. In practice, we limited the runtime to a few hours and settled for the best solution found in that time.

2.3.2 Terminology

We start with some basic definitions needed for formulating our problem.

The questions in the original questionnaire have a scale of k values of possible answers (For example, $0, \dots, 4$ in the MDS-UPDRS). Without loss of generality we renumber them $1, \dots, k$, where higher numbers indicate more severity. Each such question is translated into $k - 1$ binary features called *items*, where item i indicates that the answer to the question is at least i .

In this convention, a *visit* is a binary vector $\mathbf{v} \in \{0, 1\}^m$, where for each item i , $v_i = 1$ if and only if this item is true for that visit. We denote the j -th visit of patient p by \mathbf{v}_j^p . The sequence of visits of patient p is denoted by $(\mathbf{v}_1^p, \dots, \mathbf{v}_{m_p}^p)$, where we assume that for each patient p $m_p \geq 2$, and the visits are numbered in increasing time order. A pair of visits $(\mathbf{v}_i^p, \mathbf{v}_j^r)$ is called *proper* if $p = r$ and $i < j$. In words, the two visits should be for the same patient and they should be ordered chronologically.

The formulation assigns a *weight* $w_i \in \mathbb{R}_+$ to each item i , together forming a *weight vector* $\mathbf{w} \in \mathbb{R}_+^m$. Ensuring that item weights w_i are non-negative and defining the weight of an answer with value j as $w_1 + \dots + w_j$ guarantees that the answer weights are monotone increasing. For a visit \mathbf{v} and weights \mathbf{w} , the *score* of the visit is defined as $\mathbf{w} \cdot \mathbf{v}$.

A *longitudinal dataset* is a collection of sequences of visits, one per patient. Formally $\{(\mathbf{v}_1^p, \dots, \mathbf{v}_{m_p}^p) \mid p = 1 \dots, n\}$. For such dataset, we define S as the set of all proper pairs of visits. In other words, $S = \{(\mathbf{v}_i^p, \mathbf{v}_j^p) \mid i < j, p = 1, \dots, n\}$.

For a proper pair of visits $(\mathbf{v}_i^p, \mathbf{v}_j^p)$ and weights \mathbf{w} , if $\mathbf{w} \cdot \mathbf{v}_i^p < \mathbf{w} \cdot \mathbf{v}_j^p$ we say that the pair's order is *consistent with the weights*, or simply that the pair is *consistent*. Note the strict inequality in the last equation. If we allowed instead $\mathbf{w} \cdot \mathbf{v}_i^p \leq \mathbf{w} \cdot \mathbf{v}_j^p$, then the weight vector $\mathbf{w} = 0^m$ would be a trivial set of weights for which all proper pairs are consistent.

We are now ready to define two basic formulations of our problem.

Maximum consistency: Given a longitudinal dataset, find a weight vector \mathbf{w} that maximizes the number of consistent pairs. In other words,

$$\max |\{(\mathbf{v}_i^p, \mathbf{v}_j^p) \mid i < j \text{ and } \mathbf{w} \cdot \mathbf{v}_i^p < \mathbf{w} \cdot \mathbf{v}_j^p, p = 1, \dots, n\}|$$

.

Maximum weighted difference: Given a longitudinal dataset, find a weight vector \mathbf{w} that maximizes the weighted difference across all proper pairs. In other words,

$$\max \sum_p \sum_{i < j} (\mathbf{w} \cdot \mathbf{v}_j^p - \mathbf{w} \cdot \mathbf{v}_i^p)$$

We will introduce several variations of these objectives in the sequel, and also consider a secondary objective of **sparsity**, aiming to reduce the number of items with non-zero weights.

2.3.3 Formulations maximizing the weighted difference

Linear Programming

A basic LP formulation of the problem is

$$\begin{aligned} \max \quad & \sum_p \sum_{i < j} (\mathbf{w} \cdot \mathbf{v}_j^p - \mathbf{w} \cdot \mathbf{v}_i^p) \\ & 0 \leq w_i \leq 1 \quad i = 1, \dots, m \end{aligned} \tag{MeanDiff}$$

This problem has a closed form solution: The objective is equal to $\mathbf{w} \cdot \sum_p \sum_{i < j} (\mathbf{v}_j^p - \mathbf{v}_i^p)$. Define $\mathbf{d} = \sum_p \sum_{i < j} (\mathbf{v}_j^p - \mathbf{v}_i^p)$. Setting $w_i = 1$ if $d_i > 0$ and zero otherwise is an optimal solution.

The following formulation takes into account also the solution sparsity:

$$\max \sum_p \sum_{i < j} (\mathbf{w} \cdot \mathbf{v}_j^p - \mathbf{w} \cdot \mathbf{v}_i^p) - \gamma \sum_i w_i \tag{5}$$

$$0 \leq w_i \leq 1 \quad i = 1, \dots, m \tag{6}$$

The second term is an L1 regularization of the weights, which incentives sparsity. γ is a hyper-parameter that balances between the weighted difference objective and the aim of minimizing the number of used items. This problem too has a closed form solution, since the objective can be written as $\sum_k w_k \cdot d_k - \gamma \sum_k w_k = \sum_k w_k (d_k - \gamma)$ so setting $w_k = 1$ if $d_k - \gamma > 0$ and zero otherwise is an optimal solution.

Variable pair scores. A possible generalization of the first term in the objective is by assigning different values to different pairs: $\sum_p \sum_{i < j} q(p, i, j) (\mathbf{w} \cdot \mathbf{v}_j^p - \mathbf{w} \cdot \mathbf{v}_i^p)$. The value $q(p, i, j)$ of the pair can be used to reduce the weight of visit pairs for patients that have a lot of visits. For example, if patient p has t visits, then we can make $q(p, i, j) = \frac{1}{\binom{t}{2}}$ to give each patient equal total weight, or $q(p, i, j) = \frac{t}{\binom{t}{2}}$ to make the total weight proportional to the number of visits (as opposed to t^2). Alternatively, we can assign different weights to different pairs of visits based on their time span, as larger time gaps are expected to more strongly capture changes in disease severity.

Penalizing score drops. This approach is similar to the previous one, but instead of simply maximizing the weighted sum of differences, we would like to punish more heavily inconsistent pairs. We show this for the basic formulation. Denote by S the set of all proper visit pairs, where the elements in S are the triplets (p, i, j) such that i and j are visits of patient p with $j > i$. For each $(p, i, j) \in S$ define nonnegative variables $U_{p,i,j}$ (for up) and $D_{p,i,j}$ (down).

$$\max \sum_p \sum_{i < j} (U_{p,i,j} - \delta D_{p,i,j}) \quad (\text{MeanDiff-W})$$

$$\mathbf{w} \cdot \mathbf{v}_j^p - \mathbf{w} \cdot \mathbf{v}_i^p = U_{p,i,j} - D_{p,i,j} \quad \forall (p, i, j) \in S \quad (7)$$

$$0 \leq w_i \leq 1, \quad i = 1, \dots, m \quad (8)$$

$$U_{p,i,j}, D_{p,i,j} \geq 0 \quad (9)$$

$\delta > 1$ is the penalty coefficient for inconsistent pairs.

Claim 1. *Any optimal solution of the problem must satisfy:*

(i) *If $\mathbf{w} \cdot (\mathbf{v}_j^p - \mathbf{v}_i^p) \geq 0$, then $U_{p,i,j} = \mathbf{w} \cdot (\mathbf{v}_j^p - \mathbf{v}_i^p)$ and $D_{p,i,j} = 0$.*

(ii) If $\mathbf{w} \cdot (\mathbf{v}_j^p - \mathbf{v}_i^p) \leq 0$, then $D_{p,i,j} = -\mathbf{w} \cdot (\mathbf{v}_j^p - \mathbf{v}_i^p)$ and $U_{p,i,j} = 0$.

Proof: We prove case (i). The proof of (ii) is analogous. If $\mathbf{w} \cdot (\mathbf{v}_j^p - \mathbf{v}_i^p) = 0$ then by (7) $U_{p,i,j} = D_{p,i,j}$. The contribution of this triplet (p, i, j) to the objective is $D_{p,i,j} - \delta D_{p,i,j}$, which is negative since $\delta > 1$ unless $U_{p,i,j} = D_{p,i,j} = 0$.

If $\mathbf{w} \cdot (\mathbf{v}_j^p - \mathbf{v}_i^p) > 0$, suppose $U_{p,i,j} \neq \mathbf{w} \cdot (\mathbf{v}_j^p - \mathbf{v}_i^p)$. Define $d = U_{p,i,j} - \mathbf{w} \cdot (\mathbf{v}_j^p - \mathbf{v}_i^p)$. To satisfy (7), we get $D_{p,i,j} = d$. $d \geq 0$ due to the non-negativity constraints. Assume by contradiction that $d > 0$. The objective function then changes by:

$$U_{p,i,j} - \delta D_{p,i,j} = \mathbf{w} \cdot (\mathbf{v}_j^p - \mathbf{v}_i^p) + d - \delta d = \mathbf{w} \cdot (\mathbf{v}_j^p - \mathbf{v}_i^p) - (\delta - 1)d$$

Since $\delta > 1$ and $d > 0$, we get a strictly worse objective value than if $d = D_{p,i,j} = 0$, in contradiction to the assignment being optimal. ■

Quadratic Programming

Similarly to the LP approach, we can also introduce quadratic terms in the objective - thus formulating a quadratic programming problem.

Squaring the changes. This approach simply squares $U_{p,i,j}$ and $D_{p,i,j}$, thus giving more weight to the big changes compared to the smaller ones:

$$\max \sum_p \sum_{i < j} (U_{p,i,j}^2 - \delta D_{p,i,j}^2) \quad (10)$$

$$\mathbf{w} \cdot \mathbf{v}_j^p - \mathbf{w} \cdot \mathbf{v}_i^p = U_{p,i,j} - D_{p,i,j} \quad \forall (p, i, j) \in S \quad (11)$$

$$0 \leq w_i \leq 1 \quad i = 1, \dots, m \quad (12)$$

$$U_{p,i,j}, D_{p,i,j} \geq 0 \quad (13)$$

Penalizing drops quadratically. Instead of squaring both terms, here we do so just for the drops - so the loss from a drop is bigger than the gain from an increase of the same size. This steers the solution toward greater consistency. We do it by replacing the objective with:

$$\max \sum_p \sum_{i < j} (U_{p,i,j} - \delta D_{p,i,j}^2) \quad (\text{MeanDiff-QP})$$

Mixing linear and quadratic penalties. The caveat of the last approach is that it is tolerant to small decreases. To avoid that, we mix both linear penalty and quadratic penalties for drops, using two coefficients:

$$\max \sum_p \sum_{i < j} (U_{p,i,j} - \delta D_{p,i,j} - \delta' D_{p,i,j}^2) \quad (14)$$

In our tests we used version (**MeanDiff-QP**), as we preferred a minimal amount of hyper-parameters.

Reducing variance. Another approach utilizing quadratic programming adds a penalty for the variance of score differences. Denote by $\Delta := \frac{1}{|S|} \sum_p \sum_{i < j} (\mathbf{w} \cdot \mathbf{v}_j^p - \mathbf{w} \cdot \mathbf{v}_i^p)$ the mean difference in the progression index between pairs of visits. The new objective function is:

$$\max [\Delta - \frac{\gamma}{|S|} \sum_p \sum_{i < j} (\mathbf{w} \cdot \mathbf{v}_j^p - \mathbf{w} \cdot \mathbf{v}_i^p - \Delta)^2] \quad (\text{MeanDiff-SV})$$

Where again γ is an hyper-parameter that balances between the weighted difference objective and the objective of the differences being more stable.

2.3.4 Formulations maximizing consistency

In this section we describe formulations that aim to find integer weights that directly maximize the consistency.

Matrix representation. Recall that S is the set of all proper visit pairs, where the elements in S are the triplets (p, i, j) such that i and j are visits of patient p with $j > i$. Denote $s := |S|$. We define a matrix of differences $\mathbf{A} \in \{-1, 0, 1\}^{s \times m}$, such that for every triplet $S_l = (p, i, j)$ and item $a \in \{1, \dots, m\}$, we have $\mathbf{A}_{l,a} = (\mathbf{v}_j^p - \mathbf{v}_i^p)_a$. In words, $\mathbf{A}_{l,a}$ is the difference in the value of item a between visits j and i of patient p . Since items are binary \mathbf{A} 's entries are 1, 0 or -1.

Integer Programming (IP). We define two boolean vectors of indicators $\mathbf{I}^+, \mathbf{I}^- \in \{0, 1\}^s$, and use the following IP formulation:

$$\max \sum_{l=1, \dots, s} I_l^+ \quad (\mathbf{Cons-Int})$$

$$(\mathbf{Aw})_l \geq \varepsilon - C(1 - I_l^+) \quad l = 1, \dots, s \quad (15)$$

$$(\mathbf{Aw})_l \leq C \cdot I_l^+ \quad l = 1, \dots, s \quad (16)$$

$$(\mathbf{Aw})_l \leq -\varepsilon + C(1 - I_l^-) \quad l = 1, \dots, s \quad (17)$$

$$(\mathbf{Aw})_l \geq -C \cdot I_l^- \quad l = 1, \dots, s \quad (18)$$

$$0 \leq w_i \leq B \quad i = 1, \dots, m \text{ integer} \quad (19)$$

$$0 \leq I_l^+, I_l^- \leq 1 \quad i = 1, \dots, s \text{ integer} \quad (20)$$

Where $0 < \varepsilon \ll 1$ is a sufficiently small constant, B is an upper bound on weight values, and C is a large constant such that $C > B \cdot m + \varepsilon$. We call this problem, which maximizes consistency and requires weight integrality **Cons-Int**. The version where (19) is changed so that weights w_i can be real valued forms the mixed IP problem called **Cons**.

Lemma 1. *Let $\mathbf{w}, \mathbf{I}^+, \mathbf{I}^-$ be a feasible solution of the problem. Then for every $l = 1, \dots, s$:*

1. $(\mathbf{Aw})_l > 0$ if and only if $I_l^+ = 1$.
2. $(\mathbf{Aw})_l < 0$ if and only if $I_l^- = 1$.
3. $(\mathbf{Aw})_l = 0$ if and only if $I_l^- = I_l^+ = 0$.

Proof: (1) Assume $I_l^+ = 1$. From (15) we get $(\mathbf{Aw})_l \geq \varepsilon - C(1 - 1) = \varepsilon$. Since $0 < \varepsilon < 1$, $(\mathbf{Aw})_l > 0$. In the other direction, assume $(\mathbf{Aw})_l > 0$ and $I_l^+ = 0$. Then from (16) we get $(\mathbf{Aw})_l \leq C \cdot I_l^+ = C \cdot 0 = 0$, a contradiction. (2) Assume $I_l^- = 1$. From (17) we get $(\mathbf{Aw})_l \leq -\varepsilon + C(1 - 1) = -\varepsilon$. Since $0 < \varepsilon < 1$, $(\mathbf{Aw})_l < 0$. In the other direction, assume $(\mathbf{Aw})_l < 0$ and $I_l^- = 0$. Then from (18) we get $(\mathbf{Aw})_l \geq -C \cdot I_l^- = -C \cdot 0 = 0$, a contradiction. (3) follows from (1) and (2). ■

Claim 2. *The solution to Cons-Int maximizes consistency.*

Proof: By Lemma 1, our objective is equivalent to consistency. Denote by $\hat{\mathbf{w}}$ an integer vector of weights $0 \leq \hat{w}_k \leq B$ that achieves optimal consistency. We will show that it corresponds to a feasible solution.

For every consistent pair S_l according to $\hat{\mathbf{w}}$ we assign $I_l^+ = 1$ and $I_l^- = 0$. For every inconsistent pair S_l we assign $I_l^+ = 0$, and assign $I_l^- = 1$ if $(\mathbf{A}\hat{\mathbf{w}})_l < 0$ or $I_l^- = 0$ if $(\mathbf{A}\hat{\mathbf{w}})_l = 0$. We claim $(\hat{\mathbf{w}}, \hat{\mathbf{I}}^+, \hat{\mathbf{I}}^-)$ is a feasible solution. Constraints (19) are satisfied by assumption, and (20) by construction. Assume first S_l is consistent, i.e. $(\mathbf{A}\hat{\mathbf{w}})_l > 0$, $I_l^+ = 1, I_l^- = 0$. The constraint (15) holds since $\hat{\mathbf{w}}$ is an integer vector and $\mathbf{A} \in \{-1, 0, 1\}^{s \times m}$, so $(\mathbf{A}\hat{\mathbf{w}})_l \geq 1 > \varepsilon$. Constraint (16) holds since $(\mathbf{A}\hat{\mathbf{w}})_l \leq mB < C$. Constraint (17) holds since $(\mathbf{A}\hat{\mathbf{w}})_l \leq mB < C - \varepsilon$. Constraint (18) holds since $(\mathbf{A}\hat{\mathbf{w}})_l > 0$.

Assume now S_l is inconsistent and decreasing, i.e. $(\mathbf{A}\hat{\mathbf{w}})_l < 0$, $I_l^+ = 0, I_l^- = 1$. The constraint (15) holds since $(\mathbf{A}\hat{\mathbf{w}})_l \geq -mB > -C + \varepsilon$. (16) holds since $(\mathbf{A}\hat{\mathbf{w}})_l \leq 0$. (17) holds since $(\mathbf{A}\hat{\mathbf{w}})_l < 0$ ($(\mathbf{A}\hat{\mathbf{w}})_l \leq -1$). (18) holds since $(\mathbf{A}\hat{\mathbf{w}})_l \geq -mB > -C + \varepsilon > -C$.

Finally, assume S_l is inconsistent and unchanged, i.e. $(\mathbf{A}\hat{\mathbf{w}})_l = 0$, $I_l^+ = 0, I_l^- = 0$. The constraint (15) holds since $(\mathbf{A}\hat{\mathbf{w}})_l \geq -mB > -C + \varepsilon$. (16) holds since $(\mathbf{A}\hat{\mathbf{w}})_l \leq 0$. (17) holds since $(\mathbf{A}\hat{\mathbf{w}})_l \leq mB < C - \varepsilon$. (18) holds since $(\mathbf{A}\hat{\mathbf{w}})_l \geq 0$. ■

Sparsity. To encourage a sparse solution, we can introduce a regularization term to the objective, as before:

$$\max \sum_{l=1, \dots, s} I_l^+ - \gamma \sum_i w_i \quad (21)$$

This drives down the sum of the weights, but not sparsity per se. The following IP formulation achieves this goal. To penalize the number of non-zero weights, we introduce helper

boolean variables $\mathbf{z} \in \{0, 1\}^m$ and formulate the problem as

$$\max \sum_{l=1, \dots, s} I_l^+ - \gamma \sum_i z_i \quad (22)$$

$$(\mathbf{A}\mathbf{w})_l \geq \varepsilon - C(1 - I_l^+) \quad l = 1, \dots, s \quad (23)$$

$$(\mathbf{A}\mathbf{w})_l \leq C \cdot I_l^+ \quad l = 1, \dots, s \quad (24)$$

$$(\mathbf{A}\mathbf{w})_l \leq -\varepsilon + C(1 - I_l^-) \quad l = 1, \dots, s \quad (25)$$

$$(\mathbf{A}\mathbf{w})_l \geq -C \cdot I_l^- \quad l = 1, \dots, s \quad (26)$$

$$0 \leq w_i \leq B \cdot z_i \quad i = 1, \dots, m \text{ integer} \quad (27)$$

$$0 \leq I_l^+, I_l^- \leq 1 \quad i = 1, \dots, s \text{ integer} \quad (28)$$

$$0 \leq z_i \leq 1 \quad i = 1, \dots, m \text{ binary} \quad (29)$$

Where again $\gamma > 0$ balances between consistency and sparsity.

Claim 3. *An optimal solution of the system must satisfy $\sum_i \mathbf{z}_i = |\{i | \mathbf{w}_i > 0\}|$*

Proof: We claim $z_i = 1$ if and only if $w_i > 0$, from which the claim follows. If $w_i > 0$ then it must be that $z_i = 1$, to satisfy 27. If $z_i = 1$ but $w_i = 0$, then assigning $z_i = 0$ will increase the objective value by γ without invalidating any constraint, in contradiction to the solution's optimality. ■

Thus, the penalty term is exactly γ times the number of non-zero weights. In our tests we tried both **Cons-Int** and **Cons**, and included a regularization term for sparsity.

An alternative objective. By definition, when optimizing for consistency we do not differentiate between the cases where $\mathbf{w} \cdot (\mathbf{v}_j^p - \mathbf{v}_i^p) = 0$ and $\mathbf{w} \cdot (\mathbf{v}_j^p - \mathbf{v}_i^p) < 0$. However, no change is preferred over negative change when pursuing a monotonic score. This can be achieved using a similar formulation, by adding a penalty term for the number of strictly negative decreases, with a balance parameter γ :

$$\max \sum_{l=1, \dots, s} I_l^+ - \gamma \sum_{l=1, \dots, s} I_l^- \quad (30)$$

We note that a sparsity term can also be added to the objective.

3 Results

3.1 Comparing the performance of the different approaches

We learned the weights for each formulation and algorithm on the training set, and measured their performance in tracking disease progression, as quantified by the consistency on the left-out validation set. Table 2 shows the results when all items were used, as well as the results from applying the original weights of the complete MDS-UPDRS, its individual sections, and MoCA. The new scales outperform the MDS-UPDRS and its parts as well as MoCA, with MeanDiff-QP performing best.

Years Gap	1	2	3	4	5	6	7	8	9	10	All
Number of Pairs	417	319	242	185	134	96	71	53	38	21	1,576
MDS-UPDRS P1	49.64	56.43	55.79	64.32	62.69	68.75	70.42	73.58	73.68	76.19	58.63
MDS-UPDRS P2	53.00	58.31	60.33	69.19	72.39	78.12	83.10	90.57	92.11	95.24	64.40
MDS-UPDRS P3	50.84	54.55	47.93	51.89	55.97	54.17	69.01	75.47	81.58	80.95	54.70
MDS-UPDRS	54.44	57.05	57.44	63.24	66.42	63.54	78.87	84.91	89.47	90.48	61.48
MoCA	38.13	37.93	38.43	40.00	44.03	43.75	32.39	45.28	42.11	57.14	39.53
MeanDiff	55.40	57.68	61.57	69.73	73.88	76.04	83.10	84.91	92.11	85.71	64.85
MeanDiff-W	58.51	62.07	62.81	72.97	76.12	80.21	85.92	90.57	92.11	90.48	67.96
MeanDiff-QP	62.35	66.14	73.14	78.38	85.07	91.67	91.55	98.11	97.37	100.00	74.24
MeanDiff-SV	59.71	63.01	64.05	72.97	79.85	81.25	88.73	92.45	94.74	95.24	69.35
Cons	59.71	64.26	69.01	76.76	79.85	84.38	88.73	94.34	97.37	95.24	71.13
Cons-Int	58.27	65.52	68.18	76.76	85.07	85.42	91.55	88.68	94.74	100.00	71.32

Table 2: Performance comparison when all MDS-UPDRS and MoCA questions are used. The table shows the percentage of consistent pairs of visits for each method, for different time gaps between the visits. Time gaps are rounded to the closest year. The number in bold shows the best performer for each gap. The last column gives the weighted average percentage of consistent pairs.

Several observations can be derived from these results. First, the MoCA exhibits low consistency, likely due to two factors. Many PD patients, particularly in the early stages, do not experience significant cognitive decline. More importantly, MoCA performance is affected by practice effect, where repeated tests lead to improved scores independent of actual cog-

nitive changes [122]. Second, it is noteworthy that the simple score based on part 2 only outperforms the full MDS-UPDRS score. In particular, it outperforms part 3, which is often regarded as the most clinically relevant and reliable. This can be attributed to the influence of medications, which strongly affect the motor symptoms assessed in part 3. Changes in medication or dosage adjustments frequently lead to lower part 3 scores when patients are in the ON state (as in the dataset used here). Additionally, part 2 assessments, being self-reported, avoid the inter-rater variability that affects part 3, reducing measurement noise and improving consistency. Third, while all our new methods outperform the baseline methods, MeanDiff-QP achieves the highest consistency. Note that it outperforms both Cons and Cons-Int, which strive to directly optimize consistency, likely due to the computational hardness of the latter problems, which necessitates early stopping of the algorithms before reaching optimality.

Table 3 shows the results when only the self-reported items are used.

Years Gap	1	2	3	4	5	6	7	8	9	10	All
Number of Pairs	417	319	242	185	134	96	71	53	38	21	1,576
MDS-UPDRS P1	46.76	52.66	58.68	61.62	57.46	65.62	71.83	75.47	76.32	66.67	56.66
MDS-UPDRS P2	53.00	58.31	60.33	69.19	72.39	78.12	83.10	90.57	92.11	95.24	64.40
MDS-UPDRS	53.96	61.13	66.94	74.05	74.63	72.92	85.92	90.57	94.74	100.00	66.94
MeanDiff	54.44	61.76	66.94	72.43	76.12	73.96	84.51	92.45	94.74	100.00	67.20
MeanDiff-W	54.20	61.13	66.12	72.43	77.61	76.04	84.51	90.57	94.74	100.00	67.07
MeanDiff-QP	58.99	61.13	71.90	76.22	82.09	84.38	91.55	96.23	97.37	100.00	71.13
MeanDiff-SV	58.51	64.89	67.36	74.05	76.12	80.21	88.73	94.34	94.74	100.00	69.48
Cons	60.91	67.71	68.60	76.22	83.58	87.50	91.55	94.34	92.11	90.48	72.46
Cons-Int	54.68	64.58	67.77	74.59	80.60	88.54	90.14	92.45	92.11	100.00	69.67

Table 3: Performance comparison when only the self-reported items in MDS-UPDRS are used. See the caption of Table 2 for details. MoCA and MDS-UPDRS part 3 are excluded as they do not contain self-reported items.

When limited to self-reported items, a similar advantage of the new scale is observed. MeanDiff-QP performs on par with the Cons method, with the latter being slightly better for shorter time gaps. Note this comparison does not account for other factors such as model simplicity, which will be discussed next.

Figure 5 compared the performance of MeanDiff-QP and MDS-UPDRS when all items are used and when only self-reported items are used. Remarkably, in both cases our optimized method shows better consistency compared to MDS-UPDRS across all time gaps. Moreover, the self-reported version is almost as good as what we can get with all items.

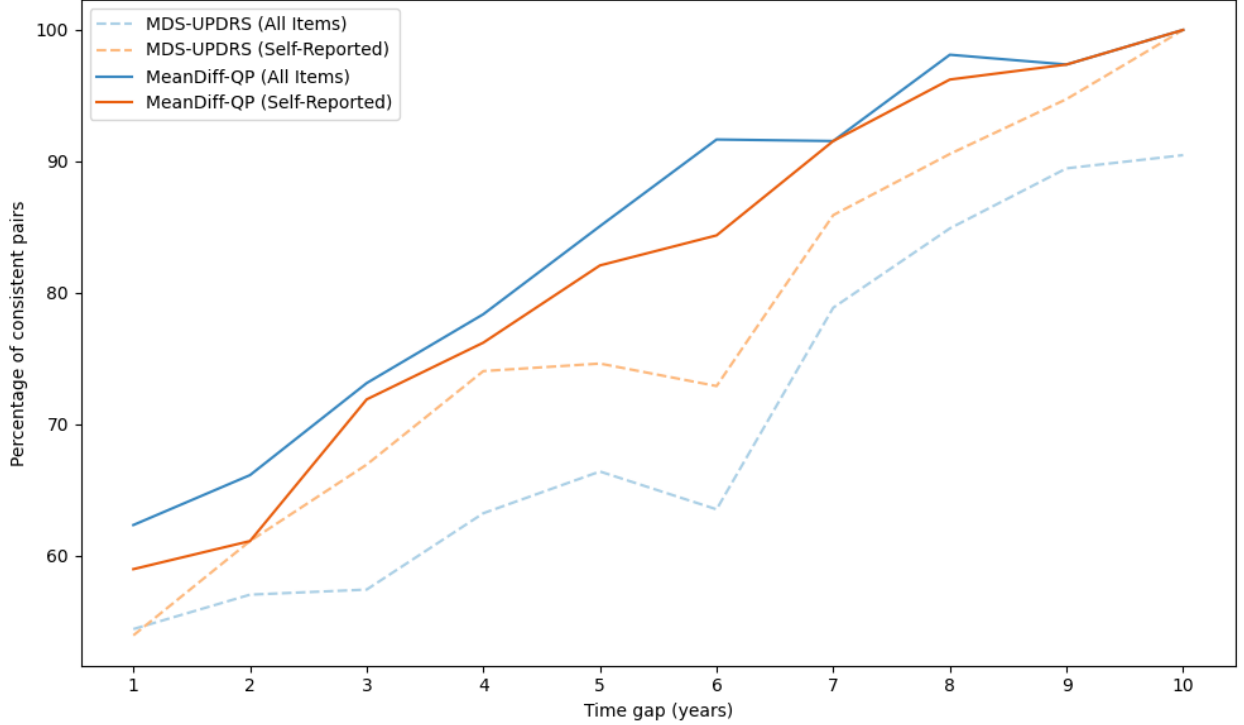


Figure 5: Percentage of consistent pairs for MDS-UPDRS and our suggested MeanDiff-QP formulation in various time gaps.

3.2 Reducing the number of items

Another potential benefit of our formulation and optimization methods can be to reduce the number of items and scoring thresholds being used, thus simplifying the scale. Figure 6 shows, for each method, its consistency and the number of non-zero items it uses. In Figure 6A all items were considered, and in Figure 6B only the self-reported items were allowed. In both cases Cons-Int achieved very good consistency, while using a very small number of items. Table 4 shows the learned weights of Cons-Int using only self-reported items. Remarkably,

only eleven questions are used, and in ten of those only one threshold value is needed. In the 11th (Getting out of bed) two thresholds are needed. Put differently, this scale uses only twelve self-reported items yet it outperforms the original 200-item MDS-UPDRS. The only scale to achieve higher consistency is Meandiff-QP with 176 items. Supplementary table S3 gives the weights of Const-Int when all items are allowed. See Supplementary 5.3 for additional information on the learned weights.

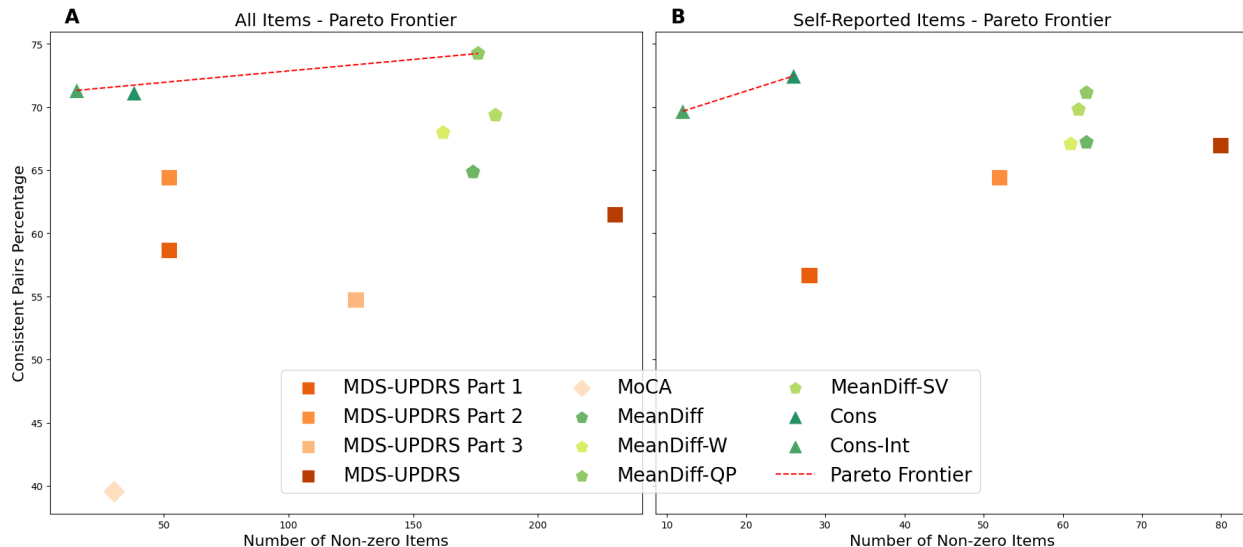


Figure 6: Consistency versus the number of items used by each method. The plots show, for each approach, the number of items it used (x axis) and the percentage of score increases between pairs of visits (y axis). A. Performance when using all items. B. Performance when using only self-reported items. The red lines indicate the Pareto optimal contour

3.3 Additional validation using PPMI

3.3.1 Initiation of symptomatic therapy

To support the clinical value of our methods, we compared their scores against two external metrics. First, we examined the relationship between a visit’s score and the time elapsed from that visit until the start of levodopa treatment. An effective scale should assign higher scores to patients who are closer to beginning treatment. Supplementary Table S4 lists the

Item	Threshold	Score
1.7 Sleep problems	1	13
1.8 Daytime sleepiness	1	25
1.10 Urinary problems	2	55
1.11 Constipation problems	1	43
1.13 Fatigue	1	30
2.1 Speech	1	44
2.2 Saliva and drooling	1	51
2.9 Turning in bed	1	66
2.11 Getting out of bed	1	52
2.11 Getting out of bed	2	45
2.12 Walking and balance	2	100
2.13 Freezing	1	67

Table 4: The scale obtained by Cons-Int using only self-reported items. Only the non zero weights are shown. The final index is obtained by summing the scores for all rows in which the item’s value is equal or larger than the threshold.

correlation between each tested method and the time to initiation of levodopa. Indeed, we see highly significant negative correlations between the scores and the time difference. Figure 7 shows the results of the best performing method in terms of the significance of correlation in each scenario: MeanDiff-QP using all items and Cons-Int using just self-reported items.

We also checked the scores concordance with the Schwab and England Activities of Daily Living (S&E ADL) scale [39]. Again, we expect a negative correlation between our disease progression score and the ADL score. Here too, the results of all scales were significant (Supplementary Table S5). Figure 8 shows the results for MeanDiff-QP, which achieved the highest correlation using all items, and was second-best using only self-reported items, surpassed only by MDS-UPDRS Part 2.

Both tests validate the relevancy of our suggested scores, showing high correlations to external data that was not a part of the training process. Many methods outperform the original

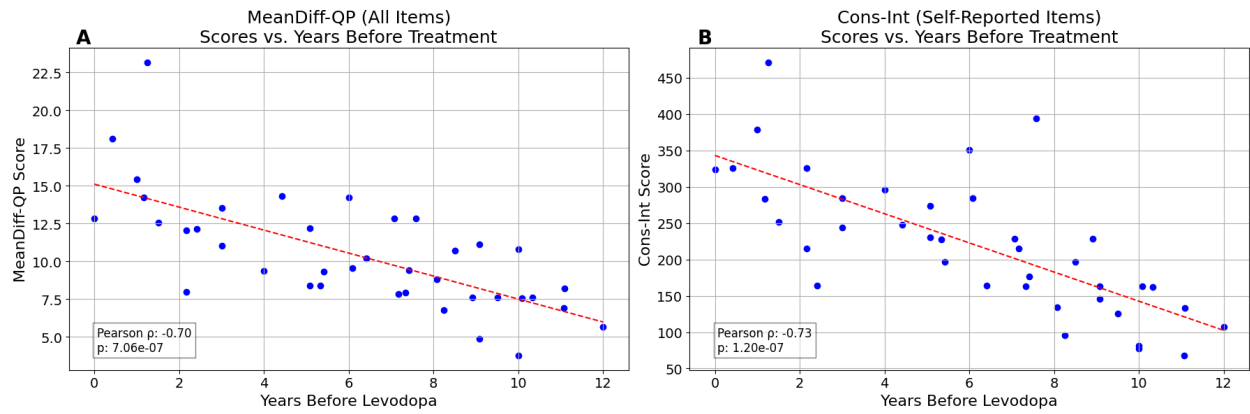


Figure 7: Total scores vs. the number of years prior to initiation of levodopa treatment. A. MeanDiff-QP using all items. B. Cons-Int using self-reported items only.

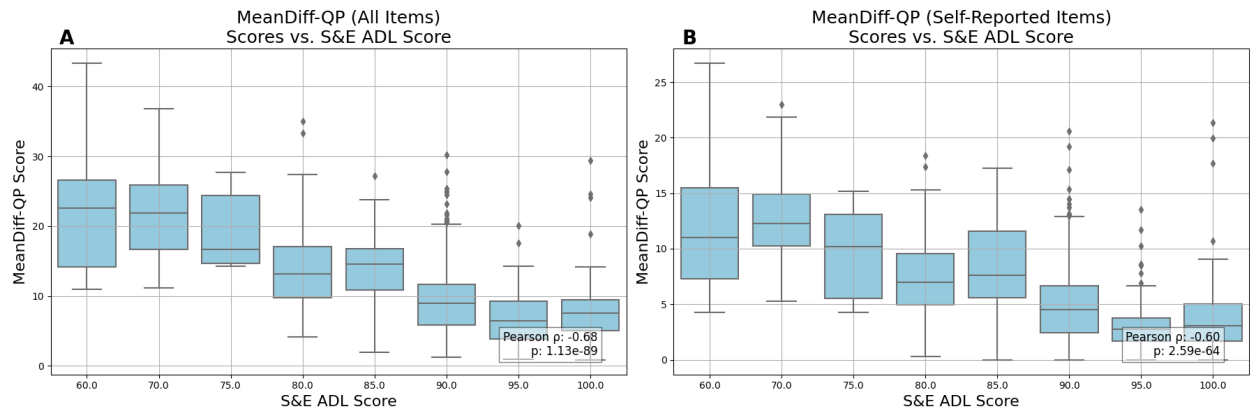


Figure 8: Total scores compared to the S&E ADL scores. A. MeanDiff-QP results when using all items. B. MeanDiff-QP results when using only self-reported items.

scales, reaching correlations of -0.73 ($p=1.20e-07$) with the **Cons-Int** method for time before levodopa treatment and -0.68 ($p=1.13e-89$) with **MeanDiff-QP** for S&E ADL.

3.3.2 Time to milestone

Recent studies suggested to quantify disease progression based on combinations of the individual item scores in the MDS-UPDRS, in addition to the total score of the scale. Brumm et al. [85] defined a set of milestones based on the MDS-UPDRS and defined as a first milestone the first time a patient reaches any of them. We wanted to see how well our progression index predicts the time it would take a patient to reach the first milestone. We tested 20 out of the 25 milestones defined in [85]; the remaining milestones were excluded due to limited availability of the necessary data. For each patient we checked which visit was the first to reach any of the milestones, and then checked the correlation between the value of our suggested progression index in each preceding visit and the time to the first milestone. A good index should exhibit a strong negative correlation, meaning that higher scores are associated with a shorter time to reaching the first milestone.

The results for methods using all items can be seen in Figure 9. All suggested scales achieved correlation below -0.39, outperforming MDS-UPDRS. The correlation coefficients and p values for all methods are available in Supplementary Table S6.

3.4 External Validation

The consistency of all the tested methods on the BeaT-PD cohort is shown in Table 5. Reassuringly, all but one method exceeded the performance of the strongest baseline scale, supporting the robustness of our approach. Validation results using only self-reported items are available in Supplementary 5.6.

The participants characteristics of the BeaT-PD cohort after applying the filtering are described in Table 6.

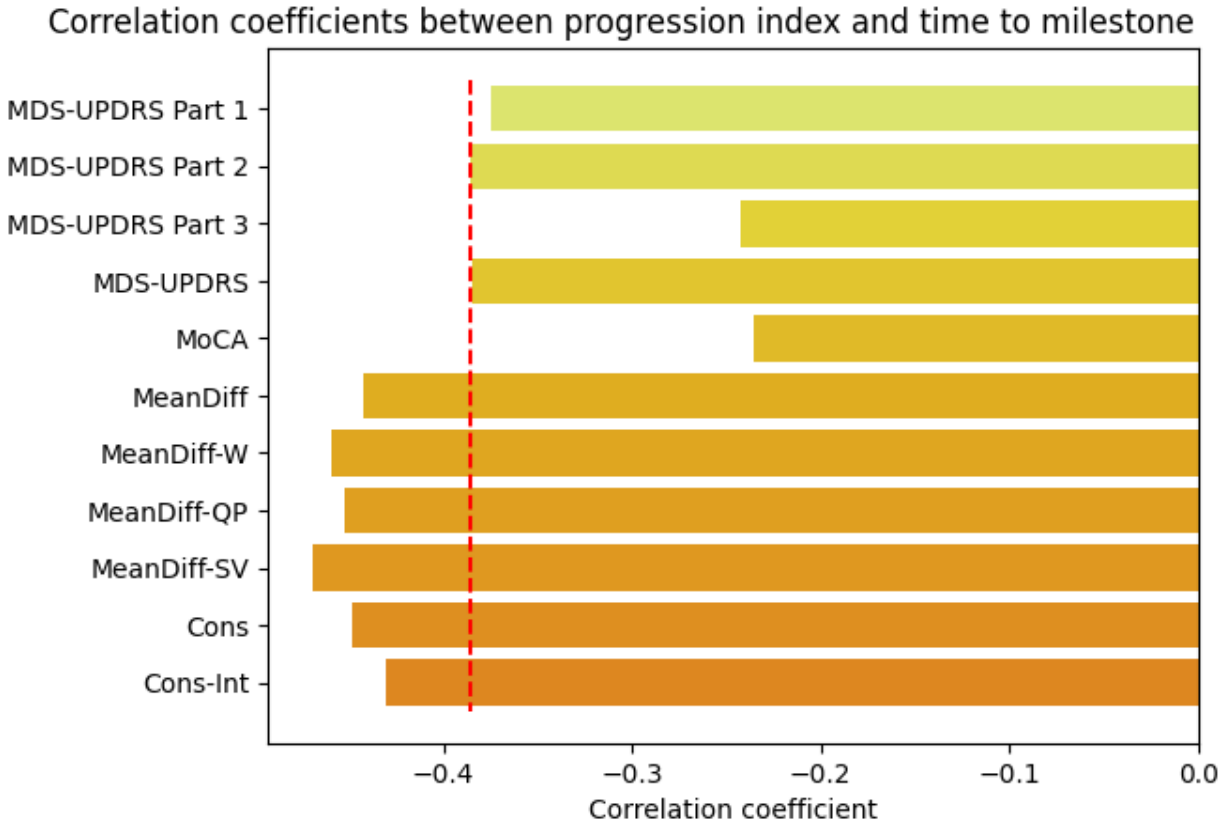


Figure 9: Correlation between the progression index and the time (in months) until the patient reaches the first milestone, as defined in [85]. All visits after the first visit where a patient reaches a milestone are excluded. The red line indicates the correlation for the MDS-UPDRS; our suggested methods exceed this threshold significantly.

Method	Consistency (%)
MDS-UPDRS P1	62.29
MDS-UPDRS P2	65.71
MDS-UPDRS P3	61.71
MDS-UPDRS	65.71
MoCA	54.29
MeanDiff	69.14
MeanDiff-W	70.29
MeanDiff-QP	67.43
MeanDiff-SV	71.43
Cons	64.57
Cons-Int	67.43

Table 5: The percentage of consistent pairs of visits for each method on the external validation BeaT-PD dataset, evaluated using the weights derived from the PPMI data.

Characteristic	Mean	Std	Min	Max
Age (years)	64.05	11.32	36	86
Disease Duration (years)	2.82	1.94	1	7
MDS-UPDRS Total Score	33.18	16.69	7	89
H&Y Stage	1.68	0.57	0	3
MoCA Total Score	24.73	3.49	17	30
Gender (M%)	77.22%			
Follow-up Time (years)	3.71	1.6	1	6
Number of Visits	2.54	0.76	2	5

Table 6: BeaT-PD participants characteristics. Statistics are shown for the cohort after filtering, consisting of 201 visits of 79 patients. The first five characteristics listed correspond to each patient’s initial visit. MDS-UPDRS: Movement Disorder Society’s Unified Parkinson’s Disease Rating Scale. MoCA: Montreal Cognitive Assessment. H&Y: Hoehn and Yahr

3.5 Hardness of the computational problem

In this section we show that the maximum consistency problem is NP-hard.

First, recall the Partial Maximum Feasible Subsystem problem (Partial Max-FS, also called Constrained Max-FS) [123, 124]. In Partial Max-FS we are given a set of linear inequality constraints, where some of them are called hard constraints and the rest are called soft constraints. We wish to find a largest cardinality subset of the constraints containing all the hard constraints and some of the soft constraints that is feasible. Partial Max-FS is NP-Hard and also hard to approximate efficiently [125]. It is also NP-hard if the variables are integer or binary, and if the inequalities are strict ($<$) [124].

Observe that the consistency maximization problem is a special case of Partial Max-FS. In our case, the non-negativity constraints are hard, and set of soft constraints is the set of examination pairs, where for each pair we want the score of the later examination to be higher. The optimal weights are those that maximize the number of constraints that are satisfied.

Theorem: The consistency maximization problem is NP-Hard, even when $B = 1$ (i.e., $\mathbf{w} \in \{0, 1\}^m$).

Proof: Observe first that any matrix $\mathbf{A} \in \{-1, 0, 1\}^{s \times m}$ can be viewed as a matrix of differences of a set of visits as defined in Section 2.3.4. This follows by forming a dataset with s patients where each patient i has exactly two visits, and the coordinate-wise differences between the item values in the two visits match the values in row \mathbf{A}_i . By the observation, we can conveniently discuss the problem in terms of constraints, where the weights are the variables.

We show a reduction from the 3-SAT problem. Given a 3-SAT instance with k variables and n clauses, we construct an instance of the consistency maximization problem as follows:

Variable Constraints: Set $m = 2 \cdot k$, where for each variable x_i , the index $2i - 1$ corresponds to the positive literal x_i , and the index $2i$ corresponds to the negation $\neg x_i$. For each variable x_i , introduce two inequalities:

$$\begin{aligned} w_{2i-1} - w_{2i} &\geq 1 \\ w_{2i} - w_{2i-1} &\geq 1 \end{aligned} \tag{31}$$

Observation: If $w_{2i-1} = 1$ and $w_{2i} = 0$, or $w_{2i-1} = 0$ and $w_{2i} = 1$, then exactly one of the inequalities (31) holds. If $w_{2i-1} = w_{2i} = 1$, or $w_{2i-1} = w_{2i} = 0$, then none holds.

Clause Constraints: For each clause C_l in the 3-SAT formula, introduce an inequality corresponding to the literals in the clause. Specifically:

$$w_i + w_j + w_k \geq 1 \tag{32}$$

where i, j, k are the indices corresponding to the literals in the clause.

Clearly, this reduction is polynomial in the size of the 3-SAT input. Note that since the variables $w_i \in \{0, 1\}$ and all the coefficients in (31) and (32) are 0, 1, or -1, the constraints ≥ 1 are equivalent to the $\geq \varepsilon$ constraints that we had in the Cons-Int formulation.

We claim that the 3-SAT instance is satisfiable if and only if there exists $\mathbf{w} \in \{0, 1\}^m$ that achieves consistency in exactly $n + k$ vectors of differences.

For proof, by the observation, out of each variable-related pair (31) at most one constraint can be satisfied by any assignment. Therefore, the maximal possible number of simultaneously feasible inequalities is $n + k$.

Assume the 3-SAT instance is satisfiable. Then there exists an assignment of the k variables that satisfies all n clauses. Construct the weight vector \mathbf{w} as follows: For each variable x_i : If x_i is True, set $w_{2i-1} = 1$ and $w_{2i} = 0$. If x_i is False, set $w_{2i-1} = 0$ and $w_{2i} = 1$. By the observation, exactly k inequalities corresponding to the variables are consistent. Since all clauses are satisfied by the assignment, each clause inequality has at least one corresponding index in \mathbf{w} set to 1, making all n clause inequalities consistent. Therefore, the total number of satisfied constraints is $n + k$.

Conversely, assume there exists a weight vector \mathbf{w} such that $n + k$ inequalities hold. Since there are k variable-related pairs, at least k of the satisfied inequalities correspond to the variable constraints. By the observation, for each variable x_i , exactly one of the inequalities (31) holds. If $w_{2i-1} = 1$ and $w_{2i} = 0$, set x_i to be True. If $w_{2i-1} = 0$ and $w_{2i} = 1$, set x_i to be False. Since the total number of satisfied inequalities is $n + k$, it follows that all n clause inequalities hold as well. By construction, a clause vector is consistent if and only if at least one of its corresponding literals is True. Therefore, the assignment satisfies all clauses. ■

3.6 Implementation details

All computations were conducted on a system with an AMD EPYC 7702 processor, featuring 128 logical CPUs (64 cores, 2 threads per core) at 2.0 GHz. The machine runs on GNU/Linux 4.15.0-65-generic within an NVIDIA DGX Server environment. Solving LP and QP problems was done using standard libraries like pulp and cvxpy. Solving IP and MIP formulations was done using the Gurobi Optimizer [126].

The first set of weights optimization formulations took each up to 30 minutes to complete using just a single thread. The second set of formulations, which aimed to maximize consistency, dealt with hard computational problem and thus were solved using all available cores and were each allotted a 24-hour time limit. Within this timeframe, an optimal solution could not be reached. However, the bound for the gap between the best solution found and the optimal solution ranged between 14.3% and 38.7% across all formulations. These values represent upper bounds, and the actual gaps are likely much smaller.

3.7 Tool and code availability

The code in this paper is available via <https://github.com/Shamir-Lab/MOPS>.

We also created an online tool that calculates the progression index using the self-reported answers, available via https://shamir-lab.github.io/MOPS/self_report_short.html.

4 Discussion

Contributions and key findings. We introduced a method for optimizing PD progression indexes by reweighting items and increments in the MDS-UPDRS and MoCA scales. The new indexes have higher precision and efficiency, benefiting both patients and clinicians. Compared to the current approach of merely summing raw item values, our scales enhance score consistency with disease progression while maintaining a simple “sum-of-items” format. Our results show that MoCA has low contribution to the scores, and most scales use MDS-UPDRS only. Notably, scaled based only on weighted self-reported items perform comparably to clinician-rated scales. Such scales enable reliance on self reports and remote monitoring. We also developed weights that substantially shorten the scale while maintaining a high level of consistency. For example, we propose a scale with eleven self-reported items and twelve weights that outperforms the original MDS-UPDRS with 59 items and 236 weights. Strong correlations with external progression markers and validation of the indexes on the external BeaT-PD cohort validate our approach.

Similar research. Several studies addressed the problem of optimal numerical encoding for ordinal categorical data [127, 128]. The choice of the encoding depends on the model used and on the size of the dataset. If the model is complex enough and there is plenty of data, usually one hot (or similar) encoding suffice [129]. Such a model can learn the optimal weights and will optimally utilize the data to achieve the desired outcome. However, in our case we aimed to construct a very simple model, where the progression index simply sums up the weights of the items. Such a requirement is important in terms of interpretability, and helps to prevent overfit.

Several previous studies dealt with optimizing the encoding of ordinal data in the context of clustering [130–132]. The Distance Learning-Based Clustering algorithm [133] heuristically finds the best numerical distances between adjacent ordinal values in a way that will make the clusters well separated. However, assuming PD severity levels are continuous and are not expected to be well clustered, these approaches are less suitable for our case.

Another approach to our application is IRT [134]. It assumes that each person’s responses are influenced by an underlying, hidden trait — in our case, PD severity — and estimates

how each item relates to this trait. While IRT has been applied to the MDS-UPDRS [135–140], it has some limitations. First, the IRT model assumptions are not fulfilled by the MDS-UPDRS [77]. In particular, the assumption that each item is measuring the same trait independently does not hold for the diverse symptoms of PD. Second, IRT can not incentivize sparsity, as it fits the optimal parameters for each item or question separately. Lastly, IRT is primarily designed for cross-sectional data and does not effectively capture changes over time, which are crucial for tracking disease progression.

Finally, researchers recently proposed re-weighting of MDS-UPDRS items by using partial least squares regression [141, 142], and have also developed distinct scores for a few different PD populations. However, unlike our work, the studies focused only on part 3, did not allow different weights to different increments of the same question, and used an internal criterion (mean to standard deviation ratio) that differs from consistency.

Early vs. advanced patients. Since PPMI mostly enrolls patients in an early stage of the disease, our data is biased towards early patients; for example, 92.8% of the exams are of patients with H&Y stage ≤ 2 . Therefore, the utility of our progression scale will be highest for earlier PD patients, and less informative for more advanced patients. While it is mathematically easy to balance the index and adjust the optimization target to give more weight to more severe patients, we decided against such a change for a few reasons. First, a progression index is much more valuable in earlier stages of the disease, since in later, more severe stages it is easier to identify the progression manifested in a wide range of symptoms. Second, giving more weight to patients with higher H&Y will introduce additional noise and bias, as these stages are characterized by specific aspects of PD, and do not capture the full range of symptoms. Moreover, the H&Y staging itself also exhibits inter-rater variability [143].

MoCA The MoCA exhibits a low consistency, and had a minimal contribution to the indexes, likely due to two factors. Many PD patients, particularly in the early stages, do not experience significant cognitive decline. More importantly, MoCA performance is affected by practice effect, where repeated tests lead to improved scores independent of actual cognitive changes [122].

Tremor. Previous research shows that the tremor items in part 3 contain limited information about the underlying state in PD and do not show worsening over time [144]. Additionally, an IRT scoring of part 3 items gives negative coefficients to the tremor items, claiming they are anti-correlative to the other part 3 items [145]. One contributing factor might be that tremor items are hard to assess accurately. Another issue might be that these items are strongly affected by PD medications like levodopa [146]. Indeed, in our computational approaches these items usually receive little or no weight, supporting the observation that they are poor indicators of PD progression.

MDS-UPDRS Part 2. It is noteworthy that the score based on part 2 only outperforms the full MDS-UPDRS score. In particular, it outperforms part 3, which is often regarded as the most clinically relevant and reliable. This can be attributed to the influence of medications, which strongly affect the motor symptoms assessed in part 3. Changes in medication or dosage adjustments frequently lead to lower part 3 scores when patients are in the ON state (as in the dataset used here). Additionally, part 2 assessments, being self-reported, avoid the inter-rater variability that affects part 3, reducing measurement noise and improving consistency. Lastly, while part 3 measures the present state, part 2 items usually ask about the last week, making them less susceptible to symptoms fluctuations.

Implications for clinical practice. Our weighted scales offer several tangible benefits for both clinicians and patients. First, by removing questions that contribute minimally to tracking PD progression, one can focus on the more meaningful indicators of progression without sacrificing diagnostic or prognostic accuracy. Second, the potential to base progression tracking on properly weighted self-reported items alone enables more frequent as well as remote evaluations, offering patients the flexibility to complete assessments at home, while reducing the burden from clinicians. Importantly, our decision to train only on data of patients in 'ON' state leads to indexes that are applicable to the real-world daily presentation of patients. Overall, the optimized index could enhance the quality and efficiency of patient care and improve long-term disease management.

Computational hardness. Our study developed two approaches to rescaling. The first optimizes a closely related but different objective than consistency. Still, the resulting scales

often show good performance. Our solutions to these formulations used polynomial algorithms. The second approach, which directly optimizes consistency, translates to an NP-hard problem (see 3.5) and we solved it by Integer Programming and Mixed Integer Programming algorithms, which take exponential time in the worst case. These algorithms could produce only near-optimal scales within the allocated time frame.

Limitations and Future Work. Our study has several limitations. First, we constructed our scales using only data from patients who are drug-naïve or in ON state. This aimed to ensure our results are applicable to patients in their typical daily conditions, where medication are not intentionally withheld. Future studies can use our methodology while focusing on more advanced PD patients who naturally experience frequent OFF-state periods. A more detailed pharmacological profile for each patient—capturing medication types, dosage, and timing—may also allow the model to re-weight items dominated by temporary symptomatic relief rather than true disease progression.

Second, due to limited computational resources, we split the data into training and test sets but did not allocate a separate validation set for extensive hyperparameter tuning. Instead, for each formulation we tried several parameter values on the training set and took one that performed best. A more systematic approach (e.g., nested cross-validation) using more computation power may lead to better parameter choices and improve the scales.

Finally, our current analysis was based on PD patients only. Incorporating data from healthy individuals can help refine the model’s specificity and sensitivity in detecting the onset of PD, especially in individuals at the borderline between healthy and prodromal state. Including prodromal patients would similarly expand the applicability of our approach, enabling earlier and more nuanced detection of progression trajectories.

PD subtypes. As mentioned earlier, PD is a very heterogeneous disease, probably an "umbrella term" for multiple disorders with similar symptoms that may have completely different underlying disease mechanisms. Recent studies used state of the art clustering approaches to separate PD patients into different clusters [147–151]. However, systematic reviews of these attempts show their results are very unreliable [152], often mistreat ordinal data as numeric, and are poorly reproducible on different datasets [153]. Should robust

subtypes be identified in the future, weight optimization could then be conducted separately for each disease subtype.

Alternative objectives. In this study we focused on consistency - the number of visit pairs where the score is strictly higher in the later visit. Other objectives can be considered, e.g., a version of consistency that prefers no change over negative change between visits. Our formulations can be adapted to this alternative, as discussed in Supplementary 2.3.4. We plan to explore such variants in future studies.

Other domains. While this study focused only on PD, the computational approach and methods provided here can lead to improvement in scales of other disease or conditions. Examples include Apgar score for newborn infants evaluation [154], the RENAL nephrology scoring system [155], the Glasgow Coma Scale [156], the Barthel Index for activities of daily living [157], the Mini-Mental State Examination for cognition [158], the NIH Stroke Scale [159] and many others. Such scores are broadly used in healthcare, and improving and simplifying them can increase their utility.

5 Supplementary

5.1 List of Abbreviations

Below is a list of all the abbreviations used in this manuscript and their meaning:

- DBS: Deep Brain Stimulation
- H&Y: Hoehn and Yahr
- ILP: Integer Linear Programming
- IRT: Item Response Theory
- LP: Linear Programming
- MDS: Movement Disorder Society
- MDS-UPDRS: Movement Disorder Society's Unified Parkinson's Disease Rating Scale
- MILP: Mixed Integer Linear Programming
- MoCA: Montreal Cognitive Assessment
- PD: Parkinson's Disease
- PPMI: Parkinson's Progression Markers Initiative
- RBD: REM sleep Behavior Disorder
- S&E ADL: Schwab and England Activities of Daily Living scale

5.2 PPMI Data

Our data (downloaded from PPMI on August 7 2024) contained information for 1,879 PD patients, 2,089 Prodromal patients and 400 Healthy Control. In our analysis we only used the PD patients' data. Supplementary Table S1 summarizes the data available for these

1,879 PD patients. Part 3 has more visits because patients are often measured twice - in ON and in OFF states. Of the 16,715 visits with Part 3 data, 7,139 were in ON state, 5,158 in OFF and 4,418 were of drug naive patients.

Exam	Total Visits	Unique Patients	Repeat Patients
MDS-UPDRS Part 1	12,453	1,492	1,284
MDS-UPDRS Part 2	12,467	1,496	1,284
MDS-UPDRS Part 3	16,715	1,497	1,285
MoCA	6,544	1,709	1,096

Table S1: Statistics on the PD patients in the PPMI. Repeat patients are patients who had more than one visit.

After removing visits in OFF-state and patients with just a single visit, we were left with 4,919 visits of 823 unique patients that contain all MDS-UPDRS parts and MoCA. Removing the baseline and screening visits of all patients left us with 4,269 visits of 763 unique patients. Finally, we filtered visits where dyskinesia interfered with the rating or where critical values were missing or misaligned. The final dataset used in the analysis consisted of 3,304 visits of 715 unique patients (an average of 4.62 visits per patient, with median time difference between consecutive visits of 1 year). Supplementary Table S2 presents the number of visits in each severity level in the final dataset, showing the bias towards early patients.

H&Y Stage	0	1	2	3	4	5
Number of Visits	16	602	2451	192	39	4

Table S2: Number of visits for each H&Y stage.

5.3 The weights learned by each approach

The learned weights for the scales based on all items, and for scales using only self-reported items, are available in the project’s launch page: <https://acgt.cs.tau.ac.il/mops/>. For each method, the values were normalized to sum to 100. Note that for MoCA items the values were flipped, so for example a threshold of 1 means 1 point below the maximal possible score.

Item	Threshold	Score
1.1 Cognitive impairment	1	16
1.7 Sleep problems	1	9
1.7 Sleep problems	2	25
2.2 Saliva and drooling	2	14
2.3 Chewing and swallowing	1	10
2.3 Eating tasks	1	13
2.8 Doing hobbies and other activities	1	7
2.9 Turning in bed	1	27
2.12 Walking and balance	2	47
2.13 Freezing	1	35
3.4b Finger tapping - Left hand	1	13
3.7b Toe tapping - Left hand	1	7
3.13 Posture	1	19
3.13 Posture	3	100
MoCA - Clock hands	< 1 (fail)	15

Table S3: The scale obtained by Cons-Int when all items can be used. Only non zero weights are shown. The index is obtained by summing the scores for all rows where the item's value is equal or larger than the threshold.

5.4 Comparison of the indexes to external scales

Method	All Items	Self-Reported Items
MDS-UPDRS Part 1	-0.31 (p=5.09e-02)	-0.25 (p=1.19e-01)
MDS-UPDRS Part 2	-0.67 (p=3.83e-06)	-0.67 (p=3.83e-06)
MDS-UPDRS Part 3	-0.46 (p=3.44e-03)	NA
MDS-UPDRS	-0.63 (p=2.07e-05)	-0.64 (p=1.15e-05)
MoCA	0.41 (p=1.04e-02)	NA
MeanDiff	-0.62 (p=2.62e-05)	-0.60 (p=4.57e-05)
MeanDiff-W	-0.64 (p=9.98e-06)	-0.61 (p=3.19e-05)
MeanDiff-QP	-0.70 (p=7.06e-07)	-0.67 (p=3.30e-06)
MeanDiff-SV	-0.69 (p=1.51e-06)	-0.61 (p=4.33e-05)
Cons	-0.62 (p=2.73e-05)	-0.68 (p=1.74e-06)
Cons-Int	-0.60 (p=6.02e-05)	-0.73 (p=1.20e-07)

Table S4: Correlation between the score of each method and the time to Levodopa. Results are shown for scores that use all items and for scores that use self-reported items only. p-values are calculated using Pearson’s ρ .

Method	All Items	Self-Reported Items
MDS-UPDRS Part 1	-0.42 (p=3.32e-29)	-0.31 (p=6.62e-16)
MDS-UPDRS Part 2	-0.62 (p=3.51e-69)	-0.62 (p=3.51e-69)
MDS-UPDRS Part 3	-0.40 (p=2.43e-26)	NA
MDS-UPDRS	-0.59 (p=1.33e-62)	-0.57 (p=2.80e-57)
MoCA	-0.41 (p=2.90e-27)	NA
MeanDiff	-0.65 (p=3.64e-78)	-0.56 (p=8.12e-54)
MeanDiff-W	-0.65 (p=2.55e-78)	-0.56 (p=8.53e-55)
MeanDiff-QP	-0.68 (p=1.13e-89)	-0.60 (p=2.59e-64)
MeanDiff-SV	-0.62 (p=4.14e-69)	-0.54 (p=1.31e-50)
Cons	-0.64 (p=7.17e-77)	-0.58 (p=1.14e-60)
Cons-Int	-0.57 (p=8.55e-58)	-0.52 (p=2.20e-45)

Table S5: Correlation between the score of each method and S&E ADL. Results are shown for scores that use all items and for scores that use self-reported items only. p-values are calculated using Pearson’s ρ .

5.5 Correlation of the indexes with the time to first milestone

Method	All Items	Self-Reported Items
MDS-UPDRS Part 1	-0.37 (p=3.75e-56)	-0.33 (p=4.54e-44)
MDS-UPDRS Part 2	-0.39 (p=9.49e-60)	-0.39 (p=9.49e-60)
MDS-UPDRS Part 3	-0.24 (p=1.39e-23)	NA
MDS-UPDRS	-0.38 (p=2.00e-59)	-0.41 (p=9.70e-68)
MoCA	-0.24 (p=2.98e-22)	NA
MeanDiff	-0.44 (p=4.69e-80)	-0.42 (p=3.84e-71)
MeanDiff-W	-0.46 (p=7.03e-87)	-0.42 (p=5.53e-72)
MeanDiff-QP	-0.45 (p=6.44e-84)	-0.42 (p=1.94e-71)
MeanDiff-SV	-0.47 (p=2.97e-91)	-0.42 (p=2.87e-71)
Cons	-0.45 (p=1.54e-82)	-0.42 (p=5.42e-73)
Cons-Int	-0.43 (p=1.80e-75)	-0.41 (p=4.55e-67)

Table S6: Correlation between each method’s score and the time to first milestone. Results are shown for scores that use all items and for scores that use self-reported items only. p-values are calculated using Pearson’s ρ .

5.6 External validation using self-reported items

Table S7 shows the consistency results on the external cohort when using only the self-reported items. In this analysis we did not filter by clinical state or presence of dyskinesia, which are relevant to part 3 of MDS-UPDRS. Only one of our suggested methods fell behind the best baseline method. Also, the best performing method using all items, MeanDiff-SV, is ranked second-best in the self-report-only setting.

Method	Consistency (%)
MDS-UPDRS P1	51.57
MDS-UPDRS P2	68.85
MDS-UPDRS	68.06
MeanDiff	69.63
MeanDiff-W	68.85
MeanDiff-QP	70.94
MeanDiff-SV	71.99
Cons	74.61
Cons-Int	66.23

Table S7: Percentage of consistent visit pairs for each method on the external validation dataset, evaluated with PPMI-derived weights based solely on self-reported items. MDS-UPDRS: Movement Disorder Society’s Unified Parkinson’s Disease Rating Scale. MoCA: Montreal Cognitive Assessment.

References

1. Ben-Shlomo, Y. *et al.* The epidemiology of Parkinson's disease. *The Lancet* **403**, 283–292 (2024).
2. Su, D. *et al.* Projections for prevalence of Parkinson's disease and its driving factors in 195 countries and territories to 2050: modelling study of Global Burden of Disease Study 2021. *BMJ* **388** (2025).
3. Parkinson, J. An essay on the shaking palsy. *The Journal of neuropsychiatry and clinical neurosciences* **14**, 223–236 (2002).
4. Parent, M. & Parent, A. Substantia nigra and Parkinson's disease: a brief history of their long and intimate relationship. *Canadian journal of neurological sciences* **37**, 313–319 (2010).
5. Pagano, G. *et al.* Age at onset and Parkinson disease phenotype. *Neurology* **86**, 1400–1407 (2016).
6. Baldereschi, M. *et al.* Parkinson's disease and parkinsonism in a longitudinal study: two-fold higher incidence in men. *Neurology* **55**, 1358–1363 (2000).
7. Shulman, L. M. Gender differences in Parkinson's disease. *Gender medicine* **4**, 8–18 (2007).
8. Zirra, A. *et al.* Gender differences in the prevalence of Parkinson's disease. *Movement disorders clinical practice* **10**, 86–93 (2023).
9. Quik, M. Smoking, nicotine and Parkinson's disease. *Trends in neurosciences* **27**, 561–568 (2004).
10. Pringsheim, T. *et al.* The prevalence of Parkinson's disease: a systematic review and meta-analysis. *Movement disorders* **29**, 1583–1590 (2014).
11. Gilks, W. P. *et al.* A common LRRK2 mutation in idiopathic Parkinson's disease. *The Lancet* **365**, 415–416 (2005).

12. Nuytemans, K. *et al.* Genetic etiology of Parkinson disease associated with mutations in the SNCA, PARK2, PINK1, PARK7, and LRRK2 genes: a mutation update. *Human mutation* **31**, 763–780 (2010).
13. Gan-Or, Z *et al.* Genotype-phenotype correlations between GBA mutations and Parkinson disease risk and onset. *Neurology* **70**, 2277–2283 (2008).
14. Deng, H., Wang, P. & Jankovic, J. The genetics of Parkinson disease. *Ageing research reviews* **42**, 72–85 (2018).
15. Farrow, S. L., Cooper, A. A. & O’Sullivan, J. M. Redefining the hypotheses driving Parkinson’s diseases research. *NPJ Parkinson’s disease* **8**, 45 (2022).
16. Cookson, M. R. & Bandmann, O. Parkinson’s disease: insights from pathways. *Human molecular genetics* **19**, R21–R27 (2010).
17. Tolosa, E. *et al.* Challenges in the diagnosis of Parkinson’s disease. *The Lancet Neurology* **20**, 385–397 (2021).
18. Postuma, R. B. *et al.* MDS clinical diagnostic criteria for Parkinson’s disease. *Movement disorders* **30**, 1591–1601 (2015).
19. McKeith, I. G. *et al.* Diagnosis and management of dementia with Lewy bodies: third report of the DLB Consortium. *Neurology* **65**, 1863–1872 (2005).
20. Wenning, G. K. *et al.* Multiple system atrophy. *The Lancet Neurology* **3**, 93–103 (2004).
21. Mahlknecht, P., Seppi, K. & Poewe, W. The concept of prodromal Parkinson’s disease. *Journal of Parkinson’s disease* **5**, 681–697 (2015).
22. Berg, D. *et al.* MDS research criteria for prodromal Parkinson’s disease. *Movement Disorders* **30**, 1600–1611 (2015).
23. Lin, Y.-Q. & Chen, S.-D. RBD: a red flag for cognitive impairment in Parkinson’s disease? *Sleep Medicine* **44**, 38–44 (2018).

24. Ponsen, M. M. *et al.* Idiopathic hyposmia as a preclinical sign of Parkinson's disease. *Annals of Neurology: Official Journal of the American Neurological Association and the Child Neurology Society* **56**, 173–181 (2004).
25. Tadaiesky, M. *et al.* Emotional, cognitive and neurochemical alterations in a premotor stage model of Parkinson's disease. *Neuroscience* **156**, 830–840 (2008).
26. Antonini, A. *et al.* The progression of non-motor symptoms in Parkinson's disease and their contribution to motor disability and quality of life. *Journal of neurology* **259**, 2621–2631 (2012).
27. Gayed, I. *et al.* The impact of DaTscan in the diagnosis of Parkinson disease. *Clinical nuclear medicine* **40**, 390–393 (2015).
28. De la Fuente-Fernández, R. Role of DaTSCAN and clinical diagnosis in Parkinson disease. *Neurology* **78**, 696–701 (2012).
29. Hoehn, M. M. & Yahr, M. D. Parkinsonism: Onset, progression and mortality. *Neurology* **17**, 427–442 (1967).
30. Alba, A. *et al.* A clinical disability rating for Parkinson patients. *Journal of Chronic Diseases* **21**, 507–522 (1968).
31. Hely, M. *et al.* Reliability of the Columbia scale for assessing signs of Parkinson's disease. *Movement disorders: official journal of the Movement Disorder Society* **8**, 466–472 (1993).
32. Fahn, S. Unified Parkinson's disease rating scale. *Recent developments in Parkinson's disease*, 153–163 (1987).
33. On Rating Scales for Parkinson's Disease, M. D. S. T. F. The unified Parkinson's disease rating scale (UPDRS): status and recommendations. *Movement Disorders* **18**, 738–750 (2003).
34. Richards, M. *et al.* Interrater reliability of the Unified Parkinson's Disease Rating Scale motor examination. *Movement Disorders* **9**, 89–91 (1994).

35. Goetz, C. G. *et al.* Movement Disorder Society-sponsored revision of the Unified Parkinson's Disease Rating Scale (MDS-UPDRS): scale presentation and clinimetric testing results. *Movement Disorders* **23**, 2129–2170 (2008).
36. Stiasny-Kolster, K. *et al.* The REM sleep behavior disorder screening questionnaire—a new diagnostic instrument. *Movement disorders* **22**, 2386–2393 (2007).
37. Chaudhuri, K. R. *et al.* The Parkinson's disease sleep scale: a new instrument for assessing sleep and nocturnal disability in Parkinson's disease. *Journal of Neurology, Neurosurgery & Psychiatry* **73**, 629–635 (2002).
38. Doty, R. L., Shaman, P. & Dann, M. Development of the University of Pennsylvania Smell Identification Test: a standardized microencapsulated test of olfactory function. *Physiology & behavior* **32**, 489–502 (1984).
39. Schwab, J. *Projection technique for evaluating surgery in Parkinson's disease* in *Third Symposium on Parkinson's Disease/E & S Livingston* (1969).
40. Nasreddine, Z. S. *et al.* The Montreal Cognitive Assessment, MoCA: a brief screening tool for mild cognitive impairment. *Journal of the American Geriatrics Society* **53**, 695–699 (2005).
41. Ciesielska, N. *et al.* Is the Montreal Cognitive Assessment (MoCA) test better suited than the Mini-Mental State Examination (MMSE) in mild cognitive impairment (MCI) detection among people aged over 60? Meta-analysis. *Psychiatr Pol* **50**, 1039–1052 (2016).
42. Calne, D. B. Treatment of Parkinson's disease. *New England Journal of Medicine* **329**, 1021–1027 (1993).
43. Lang, A. E. & Espay, A. J. Disease modification in Parkinson's disease: current approaches, challenges, and future considerations. *Movement Disorders* **33**, 660–677 (2018).
44. Group, P. S. Levodopa and the progression of Parkinson's disease. *New England Journal of Medicine* **351**, 2498–2508 (2004).

45. Shaw, K., Lees, A. & Stern, G. The impact of treatment with levodopa on Parkinson's disease. *QJM: An International Journal of Medicine* **49**, 283–293 (1980).
46. Hauser, R. A. *et al.* Extended-release carbidopa-levodopa (IPX066) compared with immediate-release carbidopa-levodopa in patients with Parkinson's disease and motor fluctuations: a phase 3 randomised, double-blind trial. *The Lancet Neurology* **12**, 346–356 (2013).
47. Nausieda, P. A. *et al.* A multicenter, open-label, sequential study comparing preferences for carbidopa-levodopa orally disintegrating tablets and conventional tablets in subjects with Parkinson's disease. *Clinical therapeutics* **27**, 58–63 (2005).
48. Olanow, C. W. *et al.* Continuous intrajejunal infusion of levodopa-carbidopa intestinal gel for patients with advanced Parkinson's disease: a randomised, controlled, double-blind, double-dummy study. *The Lancet Neurology* **13**, 141–149 (2014).
49. Bennett Jr, J. P. & Piercey, M. F. Pramipexole—a new dopamine agonist for the treatment of Parkinson's disease. *Journal of the neurological sciences* **163**, 25–31 (1999).
50. Adler, C. *et al.* Ropinirole for the treatment of early Parkinson's disease. *Neurology* **49**, 393–399 (1997).
51. Reynolds, N. A., Wellington, K. & Easthope, S. E. Rotigotine: in Parkinson's disease. *CNS drugs* **19**, 973–981 (2005).
52. Brooks, D. Dopamine agonists: their role in the treatment of Parkinson's disease. *Journal of neurology, neurosurgery & psychiatry* **68**, 685–689 (2000).
53. Tetrad, J. W. & Langston, J. W. The effect of deprenyl (selegiline) on the natural history of Parkinson's disease. *Science* **245**, 519–522 (1989).
54. Olanow, C. W. *et al.* A double-blind, delayed-start trial of rasagiline in Parkinson's disease. *New England Journal of Medicine* **361**, 1268–1278 (2009).
55. Blair, H. A. & Dhillon, S. Safinamide: a review in Parkinson's disease. *CNS drugs* **31**, 169–176 (2017).

56. Fernandez, H. H. & Chen, J. J. Monoamine oxidase-B inhibition in the treatment of Parkinson's disease. *Pharmacotherapy: The Journal of Human Pharmacology and Drug Therapy* **27**, 174S–185S (2007).
57. Schrag, A. Entacapone in the treatment of Parkinson's disease. *The Lancet Neurology* **4**, 366–370 (2005).
58. Ferreira, J. J. *et al.* Opicapone as an adjunct to levodopa in patients with Parkinson's disease and end-of-dose motor fluctuations: a randomised, double-blind, controlled trial. *The Lancet Neurology* **15**, 154–165 (2016).
59. Kaakkola, S. Clinical pharmacology, therapeutic use and potential of COMT inhibitors in Parkinson's disease. *Drugs* **59**, 1233–1250 (2000).
60. Schwab, R. S. *et al.* Amantadine in the treatment of Parkinson's disease. *Jama* **208**, 1168–1170 (1969).
61. Borovac, J. A. Side effects of a dopamine agonist therapy for Parkinson's disease: a mini-review of clinical pharmacology. *The Yale journal of biology and medicine* **89**, 37 (2016).
62. Asano, H. *et al.* Safety comparisons among monoamine oxidase inhibitors against Parkinson's disease using FDA adverse event reporting system. *Scientific Reports* **13**, 19272 (2023).
63. Marsden, C. D. & Parkes, J. D. " On-off" effects in patients with Parkinson's disease on chronic levodopa therapy. *The Lancet* **307**, 292–296 (1976).
64. Hardie, R. J., Lees, A. & Stern, G. On-off fluctuations in Parkinson's disease: a clinical and neuropharmacological study. *Brain* **107**, 487–506 (1984).
65. Quinn, N. P. Classification of fluctuations in patients with Parkinson's disease. *Neurology* **51**, S25–S29 (1998).
66. Fabbri, M., Barbosa, R. & Rascol, O. Off-time treatment options for Parkinson's disease. *Neurology and therapy* **12**, 391–424 (2023).

67. Senard, J. *et al.* Prevalence of orthostatic hypotension in Parkinson’s disease. *Journal of Neurology, Neurosurgery & Psychiatry* **63**, 584–589 (1997).
68. Stocchi, F., Vacca, L. & Radicati, F. G. How to optimize the treatment of early stage Parkinson’s disease. *Translational neurodegeneration* **4**, 1–7 (2015).
69. Rinne, U. Problems associated with long-term levodopa treatment of Parkinson’s disease. *Acta Neurologica Scandinavica* **68**, 19–26 (1983).
70. Katzenschlager, R. & Lees, A. J. Treatment of Parkinson’s disease: levodopa as the first choice. *Journal of neurology* **249**, ii19–ii24 (2002).
71. Benabid, A. L. Deep brain stimulation for Parkinson’s disease. *Current opinion in neurobiology* **13**, 696–706 (2003).
72. Vijiaratnam, N. *et al.* Exenatide once weekly over 2 years as a potential disease-modifying treatment for Parkinson’s disease: protocol for a multicentre, randomised, double blind, parallel group, placebo controlled, phase 3 trial: The ‘Exenatide-PD3’ study. *BMJ open* **11**, e047993 (2021).
73. Stevens, K. N. *et al.* Evaluation of simvastatin as a disease-modifying treatment for patients with Parkinson disease: a randomized clinical trial. *JAMA neurology* **79**, 1232–1241 (2022).
74. Patterson, C. G. *et al.* Study in Parkinson’s disease of exercise phase 3 (SPARX3): study protocol for a randomized controlled trial. *Trials* **23**, 855 (2022).
75. Gandhi, S. E. *et al.* Dopa Responsiveness in Parkinson’s Disease. *Movement Disorders Clinical Practice* **11**, 1113–1124 (2024).
76. U.S. National Library of Medicine. *ClinicalTrials.gov: Search Results for Parkinson’s Disease and MDS-UPDRS Part III* 2025.
77. Regnault, A. *et al.* Does the MDS-UPDRS provide the precision to assess progression in early Parkinson’s disease? Learnings from the Parkinson’s progression marker initiative cohort. *en. Journal of Neurology* **266**, 1927–1936 (2019).

78. Evers, L. J. *et al.* Measuring Parkinson’s disease over time: the real-world within-subject reliability of the MDS-UPDRS. *Movement Disorders* **34**, 1480–1487 (2019).
79. Horváth, K. *et al.* Minimal clinically important difference on the Motor Examination part of MDS-UPDRS. *Parkinsonism & related disorders* **21**, 1421–1426 (2015).
80. Cooley, S. A. *et al.* Longitudinal change in performance on the Montreal Cognitive Assessment in older adults. *The Clinical Neuropsychologist* **29**, 824–835 (2015).
81. Lei, L. K. *et al.* Stability of montreal cognitive assessment in individuals with mild cognitive impairment: potential influence of practice effect. *Journal of Alzheimer’s Disease* **87**, 1401–1412 (2022).
82. Aguilar-Navarro, S. G. *et al.* Validity and reliability of the Spanish version of the Montreal Cognitive Assessment (MoCA) for the detection of cognitive impairment in Mexico. *Revista Colombiana de psiquiatria (English ed.)* **47**, 237–243 (2018).
83. Freud, T. *et al.* Validation of the Russian version of the MoCA test as a cognitive screening instrument in cognitively asymptomatic older individuals and those with mild cognitive impairment. *Frontiers in medicine* **7**, 447 (2020).
84. Lifshitz, M., Dwolatzky, T. & Press, Y. Validation of the Hebrew version of the MoCA test as a screening instrument for the early detection of mild cognitive impairment in elderly individuals. *Journal of geriatric psychiatry and neurology* **25**, 155–161 (2012).
85. Brumm, M. C. *et al.* Parkinson’s Progression Markers Initiative: A Milestone-Based Strategy to Monitor PD Progression. *Neurology* (2023).
86. Dantzig, G. B. Programming in a linear structure. *Bulletin of the American Mathematical Society* **54**, 1074–1074 (1948).
87. Dantzig, G. B. Maximization of a linear function of variables subject to linear inequalities. *Activity analysis of production and allocation* **13**, 339–347 (1951).
88. Karmarkar, N., Resende, M. G. & Ramakrishnan, K. An interior point algorithm to solve computationally difficult set covering problems. *Mathematical Programming* **52**, 597–618 (1991).

89. Markowitz, H. Portfolio selection. *Journal of finance* **7**, 77–91 (1952).
90. Frank, M., Wolfe, P., *et al.* An algorithm for quadratic programming. *Naval research logistics quarterly* **3**, 95–110 (1956).
91. Sahni, S. Computationally related problems. *SIAM Journal on computing* **3**, 262–279 (1974).
92. Wong, E. *Active-set methods for quadratic programming* (University of California, San Diego, 2011).
93. Potra, F. A. & Wright, S. J. Interior-point methods. *Journal of computational and applied mathematics* **124**, 281–302 (2000).
94. Serafini, T., Zanghirati, G. & Zanni, L. Gradient projection methods for quadratic programs and applications in training support vector machines. *Optimization Methods and Software* **20**, 353–378 (2005).
95. Leung, Y. *et al.* A new gradient-based neural network for solving linear and quadratic programming problems. *IEEE Transactions on Neural Networks* **12**, 1074–1083 (2001).
96. Nazemi, A. & Nazemi, M. A gradient-based neural network method for solving strictly convex quadratic programming problems. *Cognitive Computation* **6**, 484–495 (2014).
97. Linderoth, J. A simplicial branch-and-bound algorithm for solving quadratically constrained quadratic programs. *Mathematical programming* **103**, 251–282 (2005).
98. Gomory, R. E. Solving linear programming problems in integers. *Combinatorial analysis* **10**, 25 (1960).
99. Schrijver, A. *Theory of linear and integer programming* (John Wiley & Sons, 1998).
100. Chen, D.-S., Batson, R. G. & Dang, Y. *Applied integer programming: modeling and solution* (John Wiley & Sons, 2011).
101. Tomlin, J. A. An improved branch-and-bound method for integer programming. *Operations Research* **19**, 1070–1075 (1971).

102. Gupta, O. K. & Ravindran, A. Branch and bound experiments in convex nonlinear integer programming. *Management science* **31**, 1533–1546 (1985).
103. Marchand, H. *et al.* Cutting planes in integer and mixed integer programming. *Discrete Applied Mathematics* **123**, 397–446 (2002).
104. Mitchell, J. E. Branch-and-cut algorithms for combinatorial optimization problems. *Handbook of applied optimization* **1**, 65–77 (2002).
105. Glover, F. Heuristics for integer programming using surrogate constraints. *Decision sciences* **8**, 156–166 (1977).
106. Cote, G. & Laughton, M. A. Large-scale mixed integer programming: Benders-type heuristics. *European Journal of Operational Research* **16**, 327–333 (1984).
107. Fischetti, M., Lodi, A., *et al.* Heuristics in mixed integer programming. *Wiley Encyclopedia of Operations Research and Management Science*. John Wiley & Sons, Inc, 2–23 (2010).
108. Berthold, T. *Primal heuristics for mixed integer programs* PhD thesis (Zuse Institute Berlin (ZIB), 2006).
109. Sörensen, K. & Glover, F. Metaheuristics. *Encyclopedia of operations research and management science* **62**, 960–970 (2013).
110. Rasch, G. *Probabilistic models for some intelligence and attainment tests*. (ERIC, 1993).
111. Birnbaum, A. Some latent trait models and their use in inferring an examinee’s ability. *Statistical theories of mental test scores* (1968).
112. Samejima, F. in *Handbook of item response theory* 95–107 (Chapman and Hall/CRC, 2016).
113. Muraki, E. A generalized partial credit model: Application of an EM algorithm. *Applied psychological measurement* **16**, 159–176 (1992).

114. Aitkin, M. Expectation maximization algorithm and extensions. *Handbook of item response theory* **2**, 217–236 (2016).
115. Patz, R. J. & Junker, B. W. A straightforward approach to Markov chain Monte Carlo methods for item response models. *Journal of educational and behavioral Statistics* **24**, 146–178 (1999).
116. Kim, J.-S. & Bolt, D. M. Estimating item response theory models using Markov chain Monte Carlo methods. *Educational Measurement: Issues and Practice* **26**, 38–51 (2007).
117. Marek, K. *et al.* The Parkinson’s progression markers initiative (PPMI)—establishing a PD biomarker cohort. *Annals of clinical and translational neurology* **5**, 1460–1477 (2018).
118. Movement Disorder Society. *MDS-UPDRS (Unified Parkinson’s Disease Rating Scale)* https://www.movementdisorders.org/MDS-Files1/PDFs/Rating-Scales/MDS-UPDRS_English_FINAL.
119. Roethlisberger, F. J. & Dickson, W. J. *Management and the Worker* (Harvard University Press, 1939).
120. Buckman, J. *et al.* *Thermometer Encoding: One Hot Way to Resist Adversarial Examples* in *International Conference on Learning Representations* (2018).
121. Thaler, A. *et al.* Mild cognitive impairment among LRRK2 and GBA1 patients with Parkinson’s disease. *Parkinsonism & Related Disorders* **123**, 106970 (2024).
122. Cooley, S. *et al.* Longitudinal Change in Performance on the Montreal Cognitive Assessment in Older Adults. *The Clinical neuropsychologist* **29**, 1–12 (2015).
123. Fu, Z. & Malik, S. *On Solving the Partial MAX-SAT Problem in Theory and Applications of Satisfiability Testing – SAT 2006* (eds Biere, A. & Gomes, C. P.) **4121** (Springer Berlin Heidelberg, Berlin, Heidelberg, 2006), 252–265.

124. Amaldi, E. & Kann, V. The complexity and approximability of finding maximum feasible subsystems of linear relations. *Theoretical Computer Science* **147**, 181–210 (1995).
125. Amaldi, E., Pfetsch, M. & Trotter, L. On the Maximum Feasible Subsystem Problem, IISs and IIS-Hypergraphs. *Mathematical Programming* **95** (2002).
126. Gurobi Optimization, LLC. *Gurobi Optimizer Reference Manual* 2025.
127. Potdar, K., Pardawala, T. S. & Pai, C. D. A comparative study of categorical variable encoding techniques for neural network classifiers. *International journal of computer applications* **175**, 7–9 (2017).
128. Seger, C. *An investigation of categorical variable encoding techniques in machine learning: binary versus one-hot and feature hashing* report, KTH Stocklohm (2018).
129. Poslavskaia, E. & Korolev, A. Encoding categorical data: Is there yet anything 'hotter' than one-hot encoding? *arXiv* (2023).
130. Walesiak, M. & Dudek, A. *Finding Groups in Ordinal Data: An Examination of Some Clustering Procedures in Classification as a Tool for Research* (eds Locarek-Junge, H. & Weihs, C.) (Springer Berlin Heidelberg, Berlin, Heidelberg, 2010), 185–192.
131. Johns, H. *et al.* Clustering clinical and health care processes using a novel measure of dissimilarity for variable-length sequences of ordinal states. en. *Statistical Methods in Medical Research* **29**, 3059–3075 (2020).
132. He, D. Active learning for ordinal classification based on expected cost minimization. en. *Scientific Reports* **12**, 22468 (2022).
133. Zhang, Y. & Cheung, Y.-m. An ordinal data clustering algorithm with automated distance learning. en. *Proceedings of the AAAI Conference on Artificial Intelligence* **34**, 6869–6876 (2020).
134. Rasch, G. *Probabilistic Models for Some Intelligence and Attainment Tests* (Danmarks Paedagogiske Institut, 1960).

135. Gottipati, G., Karlsson, M. O. & Plan, E. L. Modeling a Composite Score in Parkinson's Disease Using Item Response Theory. en. *The AAPS Journal* **19**, 837–845 (2017).
136. Tosin, M. H. d. S. *et al.* Item Response Theory Analysis of the MDS-UPDRS Motor Examination: Tremor vs. Nontremor Items. *Movement Disorders* **35**, 1587–1595 (2020).
137. Sheng, Y. *et al.* Modelling item scores of Unified Parkinson's Disease Rating Scale Part III for greater trial efficiency. en. *British Journal of Clinical Pharmacology* **87**, 3608–3618 (2021).
138. Luo, S. *et al.* Dissecting the Domains of Parkinson's Disease: Insights from Longitudinal Item Response Theory Modeling. en. *Movement Disorders* **37**, 1904–1914 (2022).
139. Zou, H. *et al.* Application of longitudinal item response theory models to modeling Parkinson's disease progression. en. *CPT: Pharmacometrics & Systems Pharmacology* **11**, 1382–1392 (2022).
140. Zou, H. *et al.* Increasing sensitivity in patient-reported MDS-UPDRS items for predicting medication initiation in early PD. en. *Movement Disorders Clinical Practice* **12(2)**, 148–156 (2024).
141. Dickson, S. *et al.* Re-weighting MDS-UPDRS Part II Items for Optimal Sensitivity to Parkinson's Disease Progression Using Parkinson's Progression Markers Initiative Natural History Data (P11-3.012). *Neurology* **102**, 2500 (2024).
142. L'Italien, G. *et al.* Re-weighting MDS-UPDRS Motor Items for Optimal Sensitivity to Parkinson's Disease Progression in Untreated Patients Using Parkinson's Progression Markers Initiative Data (S2.004). *Neurology* **102**, 2498 (2024).
143. Ginanneschi, A. *et al.* Evaluation of Parkinson's Disease: Reliability of Three Rating Scales. *Neuroepidemiology* **7**, 38–41 (1987).
144. Louis, E. D. *et al.* Progression of parkinsonian signs in Parkinson disease. *Archives of neurology* **56**, 334–337 (1999).

145. Tosin, M. H. d. S. *et al.* Item Response Theory Analysis of the MDS-UPDRS Motor Examination: Tremor vs. Nontremor Items. *Movement Disorders* **35**, 1587–1595 (2020).
146. Swinnen, B. E. K. S. *et al.* Tremor Is Highly Responsive to Levodopa in Advanced Parkinson’s Disease. *Movement Disorders Clinical Practice*, mdc3.14262 (2024).
147. Fereshtehnejad, S.-M. *et al.* Clinical criteria for subtyping Parkinson’s disease: biomarkers and longitudinal progression. en. *Brain* **140**, 1959–1976 (2017).
148. Lawton, M. *et al.* Developing and validating Parkinson’s disease subtypes and their motor and cognitive progression. en. *Journal of Neurology, Neurosurgery & Psychiatry* **89**, 1279–1287 (2018).
149. Zhang, X. *et al.* Data-Driven Subtyping of Parkinson’s Disease Using Longitudinal Clinical Records: A Cohort Study. en. *Scientific Reports* **9**, 797 (2019).
150. Rodriguez-Sanchez, F. *et al.* Identifying Parkinson’s disease subtypes with motor and non-motor symptoms via model-based multi-partition clustering. en. *Scientific Reports* **11**, 23645 (2021).
151. Dadu, A. *et al.* Identification and prediction of Parkinson’s disease subtypes and progression using machine learning in two cohorts. en. *npj Parkinson’s Disease* **8**, 1–12 (2022).
152. Hendricks, R. M. & Khasawneh, M. T. A Systematic Review of Parkinson’s Disease Cluster Analysis Research. *Aging and Disease* **12**, 1567–1586 (2021).
153. Mestre, T. A. *et al.* Reproducibility of data-driven Parkinson’s disease subtypes for clinical research. eng. *Parkinsonism & Related Disorders* **56**, 102–106 (2018).
154. Apgar, V. A proposal for a new method of evaluation of the newborn infant. *Curr Res Anesth Analg* **32**, 260–7 (1953).
155. Kutikov, A. & Uzzo, R. G. The R.E.N.A.L. nephrometry score: a comprehensive standardized system for quantitating renal tumor size, location and depth. *Journal of Urology* **182**, 844–853 (2009).

- 156. Teasdale, G. & Jennett, B. Assessment of coma and impaired consciousness. A practical scale. *Lancet* **2**, 81–4 (1974).
- 157. Mahoney, F. I. & Barthel, D. W. Functional Evaluation: The Barthel Index. *Md State Med J* **14**, 61–5 (1965).
- 158. Folstein, M. F., Folstein, S. E. & McHugh, P. R. "Mini-mental state": A practical method for grading the cognitive state of patients for the clinician. *J Psychiatr Res* **12**, 189–98 (1975).
- 159. Brott, T. *et al.* Measurements of acute cerebral infarction: a clinical examination scale. *Stroke* **20**, 864–870 (1989).

תקציר

מטרות

מחקר זה נועד לשפר את האופן שבו מדדי התקדמות קליניים של מחלת פרקינסון – ובפרט שאלון MDS-UPDRS (וכן מבחנים נוספים כמו MoCA לקוגניציה) – משקפים את התקדמות המחלה לאורך זמן. ציוני השאלות ב MDS-UPDRS הינם משתנים אורדינליים (בין 0 ל 4 בכל אחת מהשאלות) והשקלול שלהם כיום נעשה ע"י סכימת הציונים של כל השאלון - כאשר כל שאלה מקבלת משקל זהה. מכיוון שלא סביר להניח שלכל השאלות יש חשיבות זהה מבחינת התקדמות המחלה, וכמו כן אין סיבה להניח שמשמעות ההבדל הקליני בין כל זוג ציונים סמוכים הוא זהה, הנחתנו היתה שניתן לשפר את השימוש במידע שבשאלון.

שיטות

באמצעות בסיס נתונים גדול (PPMI) הכולל מעקב אחרי הציונים של השאלונים הנ"ל אצל חולי פרקינסון לאורך זמן, ועל ידי שימוש בשיטות של תכנון מתמטי (תכנון ליניארי, תכנון ריבועי ותכנון בשלמים) למדנו משקלים חדשים לציונים של כל שאלה המשקפים טוב יותר את התקדמות המחלה לאורך זמן. בשיטתנו שמרנו על מבנה השאלון המקורי של סכימת ציונים, שהינו נוח לשימוש ומאפשר לקלינאים הבנה של תהליך החישוב של המודל.

פונקציית המטרה שלנו באופטימיזציה היתה עקביות - הגדרנו מדד עקבי כמדד שמקבל ערך גבוה יותר בביקורים מאוחרים יותר. הפעלנו שיטות שמשפרות ישירות את העקביות (בעיה שהוכחנו שהינה NP-קשה אפילו במקרה הפשוט של משקלים בינאריים), וגם שיטות שמשפרות מדד הדומה לעקביות אבל לא זהה לו, המביאות לבעיות פתירות פולינומיאלית.

תוצאות

בבדיקה על 20% מהחולים שהמידע עליהם לא שימש לאימון, המדדים שהצענו הגיעו עד ל74% עקביות, לעומת 61% לכל היותר במדדים הקיימים. גם מדדים שמבוססים רק על שאלות בדיווח עצמי, ללא בדיקה של קלינאי, הגיעו ל71% ועקפו באופן משמעותי את המדדים הנהוגים. בשיטה של תכנון בשלמים, בתוספת רגולריזציה שמעודדת דלילות (שימוש במעט שאלות), הגענו למדד בעל עקביות מעולה תוך שימוש ב11 שאלות של דיווח עצמי בלבד. בנוסף, כאשר בחנו את הקורלציה של המדדים המוצעים למדדים נוספים שלא היו חלק מהאימון - כמו זמן עד להתחלת הטיפול בלבודופה או שאלון S&E לתפקוד יומיומי - גילינו שהיא לא יורדת מזו של המדדים הקיימים ולעיתים עולה עליה בהרבה. השיטות החישוביות שהצענו הן כלליות, ויכולות לשמש בעתיד לשיפור של מדדים קליניים בתחומים נוספים.



TEL AVIV אוניברסיטת
UNIVERSITY תל אביב

אוניברסיטת תל אביב

הפקולטה למדעים מדויקים ע"ש ריימונד ובברלי סאקלר

ביה"ס למדעי המחשב ובינה מלאכותית ע"ש בלווטניק

שקלול מחדש מיטבי של סולמות דירוג: מתודולוגיה ויישום במחלת פרקינסון

חיבור זה הוגש כעבודת גמר לתואר 'מוסמך אוניברסיטה' באוניברסיטת תל אביב
בבית הספר למדעי המחשב

על ידי

אסף בנש

בהנחיית

פרופ' רון שמיר

אוגוסט 2025