Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Genome based subtyping identifies Parkinson's disease subtypes with reproducible motor progression differences and clinical trial enrichment potential

Rafaela Ravazio
Pontifícia Universidade Católica do Rio Grande do Sul
Daniel Teixeira-dos-Santos
Universidade Federal do Rio Grande do Sul
Christian Mattjie
Pontifícia Universidade Católica do Rio Grande do Sul
Alvaro de Oliveira Franco
Universidade Federal do Rio Grande do Sul
Marco Antônio De Bastiani
Universidade Federal do Rio Grande do Sul
Andrei Bieger
Universidade Federal do Rio Grande do Sul
Eduardo R. Zimmer
Universidade Federal do Rio Grande do Sul
Joana Meneguzzo Pasquali
Pontifícia Universidade Católica do Rio Grande do Sul
Lucas S. Kupssinskü
Pontifícia Universidade Católica do Rio Grande do Sul
Rodrigo C. Barros
Pontifícia Universidade Católica do Rio Grande do Sul
Marco A. M. Schlindwein
Hospital de Clínicas de Porto Alegre
Lara Angi Souza
Universidade Federal do Rio Grande do Sul
Thomas Hugentobler Schlickmann
Universidade Federal do Rio Grande do Sul
Artur F. S. Schuh
aschuh@hcpa.edu.br

Universidade Federal do Rio Grande do Sul

Article

Keywords: precision medicine, patient stratification, genotype phenotype associations, polygenic risk scores, disease heterogeneity

Posted Date: July 10th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-10167910/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License. Read Full License

Additional Declarations: No competing interests reported.

Abstract

Parkinson's disease (PD) is clinically and biologically heterogeneous. We tested whether polygenic variation alone could identify early PD subtypes with distinct longitudinal outcomes and replication cohorts. We analyzed whole genome sequencing from 1,419 participants across the Parkinson's Progression Markers Initiative (PPMI), Parkinson's Disease Biomarkers Program (PDBP), and STEADY-PD III trial cohort (STEADY). An autoencoder trained on 1,050 nonanalytic participants learned representations from 98 PD risk and progression variants. Gaussian mixture modeling was then applied to 369 analytic participants, with the PPMI used for discovery (n = 265) and PDBP plus STEADY for validation (n = 104). Three genetically derived clusters were identified (C1, C2 and C3). C1 had the fastest and C3 the slowest MDS-UPDRS III progression in discovery (3.21 vs 2.14 points/year; p = 0.0072), a pattern replicated in validation (3.21 vs 0.49 points/year; p = 0.0034). Gait and posture progression were higher in C1, whereas rigidity progression was higher in C2, across datasets. In extended PPMI analyses, deep brain stimulation was more frequent in C1 than C2 and C3 (21.7%, 14.1%, and 13.6%), with lower age-adjusted hazards in C2 and C3 than C1 (HR 0.40, p = 0.013; HR 0.35, p = 0.0099). Clinical trial simulations suggested that enriching cohorts with C1 participants could reduce required sample size up to 45.4%. These findings support baseline genomic subtyping as a reproducible strategy to identify early PD subgroups with different motor trajectories and treatment-related milestones.

Introduction

Parkinson's disease (PD) is increasingly recognized as a heterogeneous disorder rather than a single clinicopathological entity. Patients may share core diagnostic features but differ in motor and non-motor manifestations, progression trajectories, genetic backgrounds, pathology, and treatment-related outcomes[1]. This heterogeneity makes PD a strong candidate for precision medicine and places meaningful subtyping at the center of efforts to predict progression, treatment response, and biological mechanisms[2,3].

Multiple approaches have attempted to stratify PD into more homogeneous groups. Early clinical classifications, such as tremor-dominant and postural instability gait disorder subtypes, were hypothesis-driven and useful for recognizing clinical variability, but they impose predefined categories on a multidimensional disease and may be unstable over time[4,5]. Data-driven approaches have addressed part of this limitation by allowing subtypes to emerge from multidimensional data, and recent studies have identified groups associated with distinct clinical trajectories and biomarker profiles[6,7]. However, the broader data-driven literature remains difficult to translate into practice because many studies have limited longitudinal evaluation, incomplete external validation, lack a patient-level assignment tool, and continue to rely on clinical variables that are vulnerable to medication state, assessment timing, rater variability, within-subject fluctuation, and patient-reported symptoms[2,8,9].

Genetic variation is an objective and stable input for PD subtyping, with evidence linking genetic profiles to PD risk, clinical heterogeneity, and longitudinal progression. Monogenic and risk gene studies indicate

that genetically defined forms of PD differ in age at onset, motor and cognitive progression, and other clinical features, while GWAS show that common variants contribute to the heritable risk of PD and may influence motor and cognitive outcomes[10–12]. Despite this rationale, genetic data remain underused in PD subtyping, which has relied mainly on clinical variables[2]. Therefore, we aimed to test whether clustering based solely on polygenic genomic data, independently of clinical information, could identify clinically-meaningful PD subtypes with distinct longitudinal motor progression, treatment-related outcomes, and biomarker profiles.

Results

Cohort and baseline cluster characteristics

The analytic sample included participants from PPMI (n = 265), PDBP (n = 42), and STEADY (n = 62). The PPMI, PDBP, and STEADY cohorts were comparable in terms of age, sex, years of education, race, family history, baseline MDS-UPDRS III scores, and polygenic risk scores (Table 1). Modest but statistically significant differences were observed in disease duration, which was slightly longer in PDBP than in PPMI (mean 1.0 vs 0.6 years, p = 0.001), Hoehn and Yahr stage (p = 0.017), and MDS-UPDRS I and II scores, which were highest in PDBP and lowest in STEADY (p = 0.015 and p = 0.007, respectively). More pronounced differences were noted in baseline dopaminergic therapy use, which was highest in PDBP and lowest in PPMI (97.6% vs 18.1%, p < 0.001), MoCA scores, which were lowest in PDBP and highest in STEADY (25.8 vs 27.8, p < 0.001), and the frequency of pathogenic *LRRK2* carriers, which was highest in PPMI (14.7%, p = 0.002; Table 1).

Three genetically-derived clusters based on previously described PD risk and progression SNPs were identified in both the discovery and validation cohorts (C1, C2, and C3). At baseline, the clusters were comparable in age, sex, years of education, family history, disease duration, Hoehn and Yahr stage, dopaminergic therapy use, MDS-UPDRS I total score, MoCA score, risk PRS, and progression PRS. In PPMI, race differed across clusters, with white race ranging from 90.2% in C1 to 100% in C3 (p = 0.012), but this difference was not observed in PDBP plus STEADY. Pathogenic variants in *LRRK2*, *SNCA*, and *PRKN* did not differ across clusters. Pathogenic *GBA1* variants differed only between PDBP and STEADY, being present in C1 and C2 but absent in C3 (15.4%, 13.5%, and 0%, respectively; p = 0.039; Table 2).

Motor progression differences between clusters

MDS-UPDRS II scores did not differ between clusters at baseline, year 3, or in yearly progression in either cohort. MDS-UPDRS III scores did not differ at baseline in either cohort, but differed at year 3 in both datasets, with higher scores in C1 and lower scores in C3. In PPMI, year 3 MDS-UPDRS III scores were 31.0 in C1 and 26.0 in C3 (overall p = 0.045; C1 vs C3 p = 0.015). In PDBP plus STEADY, the corresponding values were 30.0 and 22.4 (overall p = 0.036; C1 vs C3 p = 0.008; Table 3 and Fig. 2). Yearly MDS-UPDRS III progression also differed across clusters in both cohorts, ranging from 3.21

points/year in C1 to 2.14 in C3 in PPMI (overall p = 0.0072; C1 vs C3 p = 0.0203), and from 3.21 in C1 to 0.49 in C3 in PDBP plus STEADY (overall p = 0.0034; C1 vs C3 p = 0.0008; Table 3 and Fig. 2).

In motor subscore analyses, gait and posture and rigidity were the only domains with yearly progression differences in both cohorts. Gait and posture progression was higher in C1 than in C3 in PPMI (0.54 vs 0.39 points/year, p = 0.020), and higher in C1 than in C2 and C3 in PDBP plus STEADY (0.85 vs 0.09 and 0.23 points/year, both p < 0.0001). Rigidity progression also differed between the cohorts, mainly due to greater progression in C2 than in C3 in PPMI (0.64 vs 0.37 points/year, p = 0.001) and in PDBP plus STEADY (0.48 vs − 0.13 points/year, p = 0.003). Tremor, bradykinesia, and speech scores did not show consistent differences across datasets (Supplementary Table 2).

Long-term DBS and treatment-related outcomes

In PPMI, over up to 10 years of follow-up, DBS occurred in 20 of 92 C1 participants, 13 of 92 C2 participants, and 11 of 81 C3 participants, corresponding to crude event rates of 21.7%, 14.1%, and 13.6%, respectively. Kaplan-Meier analysis was not significant (p = 0.2402; Fig. 3). However, in Cox regression adjusted for baseline age, C2 and C3 had lower hazard ratios for DBS surgery than C1: HR 0.40, 95% CI 0.19 to 0.82, p = 0.013, and HR 0.35, 95% CI 0.16 to 0.78, p = 0.0099, respectively. Older baseline age was also associated with lower DBS hazard, HR per 10 years 0.41, 95% CI 0.30 to 0.55, p < 0.0001 (Supplementary Table 3).

In PPMI, longitudinal treatment analyses over up to 10 years showed no overall LEDD differences between clusters (Supplementary Table 4). LEDD increased in all clusters, with estimated yearly slopes of 93.5 units/year in C1, 70.1 in C2, and 86.9 in C3. Compared with C1, C2 had a lower yearly LEDD increase (β = −23.49 units/year, 95% CI − 38.16 to − 8.81, p = 0.0017; Supplementary Table 5). There were no overall or trajectory differences in levodopa response between clusters (Supplementary Tables 6 and 7).

Biomarker characterization

Baseline analyses showed no adjusted cluster differences for CSF Aβ1–42, CSF alpha synuclein, CSF hemoglobin, GFAP, CSF NfL, serum NfL, CSF a-synuclein, pTau, or tTau (Supplementary Table 8). SAA positivity was numerically higher in C1 and C2 than in C3, but did not reach statistical significance (92.0%, 90.9%, and 80.8%, respectively; overall p = 0.052; C1 vs C3 p = 0.060). Among SAA kinetic measures, the mean time to maximum slope at 24 h was higher in C1 than in C2 and C3 (11.6 h vs 9.7 h and 9.6 h; overall p = 0.026). Other SAA kinetic measures did not differ across clusters (Supplementary Table 9).

Clinical trial enrichment calculations

In the prespecified enrichment analysis, the original cohort distribution, in which C1 represented 32% of participants, required 388 participants per arm. Increasing the C1 fraction to 50%, 75%, and 100% reduced the required sample size to 329, 264, and 212 participants per arm, corresponding to reductions of 15.2%, 32%, and 45.4%, respectively (Fig. 4 and Supplementary Table 10).

Discussion

In this study, we used genomic data to identify early PD subtypes associated with longitudinal motor progression across independent cohorts. Three genetically-derived clusters were identified (C1, C2, and C3), with C1 showing the most consistent pattern of faster motor progression over the first three years. This finding was observed in the PPMI discovery cohort and replicated in the PDBP/STEADY external validation cohort, with higher yearly MDS-UPDRS III progression in C1 and lower progression in C3, accompanied by parallel differences in gait, posture, and rigidity progression. These motor progression findings were not explained by baseline demographic or clinical differences unrelated to the subtyping process. Longer-term analyses, available in PPMI with up to 10 years of follow-up, showed a higher cumulative incidence of DBS surgery in C1, with lower age-adjusted DBS hazard ratios in C2 and C3 than in C1, whereas C2 showed a lower yearly LEDD increase than C1. Together, these findings show that SNP-based subtype assignment can identify early PD patients with different motor progression trajectories.

By using polygenic SNP information as the primary input, with no clinical data, our study tested whether baseline genetic variation could identify subgroups that later diverged in motor progression. This provides a different perspective from most data-driven longitudinal studies, which have defined PD progression subtypes based on clinical variables or clinical trajectories[2,6,7]. Clinical data remain essential for determining whether a subtype is meaningful, and our own *post hoc* analysis relied on longitudinal motor outcomes. However, clinical scales can be influenced by medication status, disease stage at assessment, inter-rater variability, within-subject fluctuations, and patient-reported symptoms[8,9]. Genomics data offer a fixed, objectively measured, baseline input that can be applied before longitudinal clinical divergence becomes evident. In this context, the motor progression differences observed in our study suggest that polygenic information derived from known PD risk and progression SNPs can support a genetics-driven approach to PD subtype discovery.

The replication of the motor progression pattern in PDBP/STEADY addresses a central limitation in PD subtyping: external validation remains uncommon, and other data-driven studies have identified subtypes without independent cohort testing[23–25], a limitation also emphasized in recent critical and purpose-driven frameworks for PD subtyping[2,3]. Even among validated or multicohort studies, interpretation may remain constrained by baseline comparability, outcome harmonization, and treatment state[6,26,27]. In several of these studies, faster-progressing or clinically distinct subtypes already differed at baseline in age at diagnosis, sex, Hoehn and Yahr stage, motor phenotype, treatment exposure, LEDD, baseline motor severity, cognition, sleep, autonomic measures, or age at onset, depending on the

study[7,26,27]. These differences make it difficult to determine whether longitudinal divergence reflects a reproducible subtype signal or pre-existing clinical and demographic structure.

Medication state adds another source of possible heterogeneity, because longitudinal MDS-UPDRS III assessments should ideally be compared under similar medication conditions to reduce treatment-related bias[28]. However, several longitudinal subtyping studies did not explicitly require same state MDS-UPDRS III assessments at baseline and follow up, or addressed treatment effects indirectly through several different means[6,7,26,27]. In our study, subtype assignment was based on genomic data, motor outcomes were evaluated downstream, and analyses were restricted to early PD with harmonized 3 year follow up, same state MDS-UPDRS III assessments in either cohort. This study design makes it less likely that the motor progression pattern reflects baseline severity, treatment state, or cohort composition alone.

We identified a higher cumulative incidence of DBS in C1 than in C2 and C3 in Cox proportional hazards, baseline-age-adjusted models. Previous genetic studies of DBS in PD have focused mainly on whether monogenic variants modify post-surgical outcomes, particularly in *LRRK2*, *GBA1*, and *PRKN*-associated disease[29–32]. The polygenic background has only recently been examined as a potential modifier of DBS response[33]. Our findings address a different question: whether common-variant-based genomic structure can identify patients more likely to reach an advanced treatment milestone, rather than modifying outcomes after DBS has already been performed. This distinction is relevant because purpose-driven PD subtyping has been proposed as a means of prognostication and treatment selection, and DBS studies have emphasized that improved prediction could support candidate selection for a treatment whose response varies across patients[3,33]. Thus, although DBS is influenced by clinical eligibility, referral patterns, and local practice, the higher incidence of DBS in C1 suggests that genetically derived subtypes may reflect a trajectory of earlier progression toward advanced therapy. This finding should be interpreted cautiously because DBS was available only in PPMI, but its consistency with the motor progression results supports its role as an additional treatment milestone associated with C1.

Purpose-driven frameworks for PD subtyping emphasize evaluating subtypes based on their ability to generate clinically useful insights, including prognostication, treatment selection, and trial enrichment[3]. In this context, our clinical trial enrichment analysis addresses one of the main translational goals of PD subtyping. Heterogeneity may reduce statistical power in clinical trials, particularly for disease-modifying interventions, and can obscure treatment effects if only a subset of patients is biologically or clinically likely to progress during follow-up. This issue has been emphasized in precision medicine frameworks for PD and is illustrated by the negative primary results of two high-quality α-synuclein antibody trials despite a strong biological rationale[34–36]. Our enrichment simulation showed that increasing the proportion of C1 participants could reduce the required sample size by 15.2%, 32%, and 45.4% under 50%, 75%, and 100% enrichment scenarios, respectively. These estimates are theoretical, but they show that subtype assignment may have practical value beyond clinical prognostication.

Conventional blood and CSF biomarkers showed no adjusted differences across clusters, including Aβ1–42, tau, NfL, GFAP, and total α-synuclein. The SAA results were more ambiguous. SAA positivity was higher in C1 and C2 than in C3, with p values close to statistical significance, while the mean time to maximum slope at 24 h differed across clusters. Prior studies suggest that SAA positivity captures α-synuclein seeding biology, and that kinetic parameters may provide additional information about disease heterogeneity, particularly cognitive decline. However, recent PPMI analyses did not find robust associations between baseline SAA status or kinetic parameters and manifest PD progression[37,38]. Therefore, these findings should be interpreted as exploratory. They suggest a possible relationship between genomic subtype and α-synuclein seeding dynamics, but limited power and assay variability remain plausible explanations.

This study has limitations. The model was developed using SNPs identified largely in European-ancestry populations and validated primarily in cohorts of European ancestry, potentially limiting generalizability to more diverse populations[39]. It was also restricted to individuals with early PD, and its applicability to later disease stages remains uncertain. Although externally validated, the sample size was modest, and the model captured only a subset of PD genetic architecture, excluding rare variants, structural variants, and gene–environment interactions. Cluster selection was optimized for motor burden and progression, and therefore may not capture other dimensions of PD heterogeneity. Finally, several secondary analyses were exploratory, and the clinical utility of the model will require prospective validation in independent cohorts.

Despite these limitations, the study has several strengths. It used biological input data rather than clinical features for subtype discovery, reduced major demographic and clinical confounding at baseline, and validated motor progression patterns in an independent cohort. The analytic design also minimized information leakage by training the autoencoder in a non-overlapping set of WGS participants who were not included in the discovery or validation analyses. In addition, the study provides a practical algorithm for assigning individual subtypes based on genomic data. In clinical research, this approach may help identify patients who are more likely to progress over short follow-up intervals, thereby improving trial efficiency. For clinical practice, the algorithm provides an externally validated and deployable framework for assigning individual patients to genetically derived subtypes.

In conclusion, our findings show that baseline genomic information can identify early PD subtypes with reproducible differences in motor progression across independent cohorts. Using an SNP-based autoencoder and GMM framework, we assigned individual patients to genetically derived clusters that differed in longitudinal motor trajectories and advanced treatment-related outcomes. These results suggest that common variant structure can contribute to clinically-meaningful stratification before longitudinal clinical divergence becomes evident. The approach offers an externally validated, deployable subtype-assignment tool with a defined purpose: motor prognostic stratification. If replicated in broader ancestries, disease stages, and prospective settings, this framework may support trial enrichment, prognostic modeling, and future genotype-informed clinical workflows.

Methods

Study design

This observational study used data from three longitudinal PD cohorts: the Parkinson's Progression Markers Initiative (PPMI), the Parkinson's Disease Biomarkers Program (PDBP), and the STEADY-PD III trial cohort (STEADY). PPMI was used as the discovery cohort, whereas PDBP and STEADY were combined as an external validation cohort. We applied an unsupervised machine learning framework to PD-related single-nucleotide polymorphisms (SNPs) to identify clusters associated with longitudinal motor burden and progression. The clusters were then characterized using extended clinical outcomes, biomarkers, and clinical trial enrichment simulations. The overall workflow is shown in Fig. 1.

Setting and Participants

Participants were selected from PPMI and from AMP PD for PDBP and STEADY according to prespecified criteria: (1) PD diagnosis according to each cohort protocol, (2) available whole-genome sequencing, (3) early disease stage at enrollment, defined as disease duration ≤ 2 years in PPMI and PDBP and ≤ 3 years in STEADY, and (4) MDS-UPDRS motor assessments at baseline and year 3. To improve cohort comparability, disease duration was restricted to up to 2 years in PPMI and PDBP and up to 3 years in STEADY. Baseline and year 3 MDS-UPDRS assessments were also required to be performed in the same medication state, either both OFF or both ON. The analytic sample included 265 PPMI, 42 PDBP, and 62 STEADY participants. Intermediate visits between baseline and year 3 were retained for longitudinal models. Detailed cohort descriptions have been published previously[13–15].

This secondary analysis used de-identified data from PPMI, PDBP, and STEADY-PD III, accessed through PPMI and AMP-PD according to the applicable data use terms. The original cohorts were approved by the Institutional Review Board, Independent Ethics Committee, or Research Ethics Board at each participating site and were conducted in accordance with the Declaration of Helsinki and applicable local regulations. Written informed consent was obtained from all participants for study participation. Additional information on participating sites and study governance is available through the PPMI clinical sites page (https://www.ppmi-info.org/about-ppmi/ppmi-clinical-sites), the PDBP project documentation (https://pdbp.ninds.nih.gov/projects-we-support), and the STEADY-PD III ClinicalTrials.gov record (https://clinicaltrials.gov/study/NCT02168842).

Genetic data extraction

Whole-genome sequencing (WGS) data were obtained from AMP-PD release 4. We selected 98 SNPs previously associated with PD susceptibility or progression in European populations, including 90 risk-associated variants[10] and 8 variants associated with motor, cognitive, or non-motor progression in prior studies[11,16,17]. Genotypes were encoded additively as 0, 1, or 2, depending on the number of alternate alleles. The full list of included variants is provided in Supplementary Table 1. After cluster assignment,

we calculated polygenic risk scores using published effect sizes and extracted pathogenic or likely pathogenic variants in *LRRK2*, *GBA1*, *SNCA*, and *PRKN*.

Clinical variables

For all cohorts, we extracted baseline age, sex, race, years of education, PD family history, disease duration, Hoehn and Yahr stage[18], dopaminergic therapy use, MDS-UPDRS Parts I to III[19], and MoCA scores[20]. Longitudinal MDS-UPDRS Parts II and III and MoCA data were extracted from baseline through year 3. From MDS-UPDRS Parts II and III, we derived tremor, bradykinesia, rigidity, gait and posture, and speech subscores, as described in the Supplementary Data.

Additional PPMI analyses used up to 10 years of extended phenotyping data, including levodopa equivalent daily dose (LEDD)[21], levodopa challenge assessments, and deep brain stimulation (DBS) surgery, as well as blood and cerebrospinal fluid (CSF) biomarkers and seed amplification assay (SAA) positivity and reaction kinetic measures[22].

Unsupervised machine learning framework

An autoencoder was trained to learn a compact representation of the selected PD-related SNPs. To prevent information leakage, the model was trained exclusively on 1,050 PD participants from PPMI, PDBP, and STEADY who had WGS data but did not meet the remaining analytic inclusion criteria. Latent vectors from the PPMI discovery cohort were then clustered using a Gaussian Mixture Model. The fitted model was frozen and applied without modification to PDBP and STEADY.

Candidate solutions were filtered for stability and interpretability. Partitions with minimal patient reassignment were treated as equivalent using a 5% Adjusted Rand Index threshold, and retained clusters were required to include at least 15% of participants in both discovery and validation cohorts. Following a purpose-driven approach[3], we prioritized solutions associated with MDS-UPDRS III burden or progression in both cohorts and selected the solution with the most consistent motor separation.

Statistical and post-hoc analyses

Cluster comparisons used ANOVA, Kruskal-Wallis, chi-square, t-tests, or Mann-Whitney U tests, depending on variable type and distribution. Primary longitudinal outcomes were MDS-UPDRS II and III total scores, evaluated at baseline, year 3, and as annual progression from baseline to year 3. Annual progression was estimated using linear mixed-effects models that included cluster, follow-up time, and the cluster-by-time interaction as fixed effects, with participant-level random intercepts. Cluster contribution was tested with likelihood ratio tests comparing models with and without cluster terms. The same approach was applied to MDS-UPDRS motor subscores.

Additional PPMI analyses evaluated LEDD, levodopa response, DBS, biomarkers, SAA, and clinical trial enrichment. LEDD and levodopa response were analyzed over up to 10 years using linear mixed-effects models. DBS was analyzed using Kaplan Meier curves, log-rank tests, and Cox models adjusted for baseline age. Baseline biomarker differences were evaluated using ANCOVA adjusted for age, sex,

weight, and height. SAA positivity was compared with chi-square testing, and SAA kinetic measures were compared using ANOVA or Kruskal-Wallis tests. Clinical trial enrichment was simulated for a two-arm randomized trial using annual MDS-UPDRS III progression, alpha 0.05, 80% power, one-year follow-up, and a 30% treatment effect. Analyses were performed in Python 3.11.13, with statistical significance defined as $p \leq 0.05$.

Declarations

Competing interests

All authors declare no financial or non-financial competing interests.

Author Contribution

Conceptualization: D.T.d.S., R.R., C.M., R.C.B., E.Z., A.F.S.S. Methodology: D.T.d.S., R.R., C.M., M.A.D.B., R.C.B., E.Z., A.F.S.S. Software: D.T.d.S., R.R., C.M. Validation: D.T.d.S., R.R., C.M. Formal analysis: D.T.d.S., R.R., C.M., M.A.D.B. Investigation: D.T.d.S., R.R., C.M., M.A.D.B. Resources: Á.d.O.F., A.B., G.M.P., J.M.P., J.R.C., L.S.K., L.V.M., M.A.M.S., L.A.S., T.H.S., S.D.P. Data Curation: D.T.d.S., R.R., C.M. Writing—Original Draft: D.T.d.S., R.R., C.M. Writing—Review & Editing: D.T.d.S., R.R., C.M., Á.d.O.F., A.B., G.M.P., J.M.P., J.R.C., L.S.K., L.V.M., M.A.D.B., M.A.M.S., L.A.S., T.H.S., S.D.P, M.B.K, S.B.C. Visualization: D.T.d.S., R.R., C.M. All authors read and approved the final manuscript.

Acknowledgement

This work was supported by The Michael J. Fox Foundation for Parkinson's Research through the Data Driven Subtyping and Stratification Program, Grant ID MJFF 023158. For the purpose of open access, the author has applied for a CC BY public copyright license to all Author Accepted manuscripts arising from this submission.

Data Availability

Data used in the preparation of this article were obtained from the Parkinson's Progression Markers Initiative (PPMI; https://www.ppmi-info.org), the Parkinson's Disease Biomarkers Program (PDBP; https://pdbp.ninds.nih.gov/), and the STEADY-PD3 cohort available through the Accelerating Medicines Partnership Parkinson's Disease platform (AMP-PD; https://amp-pd.org/unified-cohorts/steady-pd-3).

Code availability

Codes used to extract relevant clinical in PPMI data such as DBS candidacy and treatment response measures are available at: https://github.com/MJFF-ResearchCommunity/Useful-PPMI-Clinical-Codes. The trained autoencoder, fitted Gaussian Mixture Model, and code for individual patient subtype assignment from genomic data are publicly available at https://github.com/Malta-Lab/SNP2PDsubtype. Additional methodological details, including model architecture, hyperparameter grids, stopping criteria, and post hoc selection rules, are available at https://zenodo.org/records/20931762.

References

1. Wüllner, U. *et al.* The heterogeneity of Parkinson's disease. *J. Neural Transm.* 130, 827–838 (2023).
2. Mestre, T. A. *et al.* Parkinson's Disease Subtypes: Critical Appraisal and Recommendations. *J. Parkinsons. Dis.* 11, 395–404 (2021).
3. Marras, C. *et al.* Transitioning from Subtyping to Precision Medicine in Parkinson's Disease: A Purpose-Driven Approach. *Mov. Disord.* 39, 462–471 (2024).
4. Zetusky, W. J., Jankovic, J. & Pirozzolo, F. J. The heterogeneity of Parkinson's disease: clinical and prognostic implications. *Neurology* 35, 522–526 (1985).
5. Simuni, T. *et al.* How stable are Parkinson's disease subtypes in de novo patients: Analysis of the PPMI cohort? *Parkinsonism Relat. Disord.* 28, 62–67 (2016).
6. Dadu, A. *et al.* Identification and prediction of Parkinson's disease subtypes and progression using machine learning in two cohorts. *NPJ Parkinsons Dis* 8, 172 (2022).
7. Hähnel, T. *et al.* Progression subtypes in Parkinson's disease identified by a data-driven multi cohort analysis. *NPJ Parkinsons Dis* 10, 95 (2024).
8. Evers, L. J. W., Krijthe, J. H., Meinders, M. J., Bloem, B. R. & Heskes, T. M. Measuring Parkinson's disease over time: The real-world within-subject reliability of the MDS-UPDRS. *Mov. Disord.* 34, 1480–1487 (2019).
9. AlMahadin, G. *et al.* Parkinson's disease: current assessment methods and wearable devices for evaluation of movement disorder motor symptoms - a patient and healthcare professional perspective. *BMC Neurol.* 20, 419 (2020).
10. Nalls, M. A. *et al.* Identification of novel risk loci, causal insights, and heritable risk for Parkinson's disease: a meta-analysis of genome-wide association studies. *Lancet Neurol* 18, 1091–1102 (2019).
11. Tan, M. M. X. *et al.* Genome-Wide Association Studies of Cognitive and Motor Progression in Parkinson's Disease. *Mov Disord* 36, 424–433 (2021).
12. Smith, L. J., Lee, C.-Y., Menozzi, E. & Schapira, A. H. V. Genetic variations in GBA1 and LRRK2 genes: Biochemical and clinical consequences in Parkinson disease. *Front. Neurol.* 13, 971252 (2022).
13. Parkinson Progression Marker Initiative. The Parkinson Progression Marker Initiative (PPMI). *Prog. Neurobiol.* 95, 629–635 (2011).
14. Rosenthal, L. S. *et al.* The NINDS Parkinson's disease biomarkers program. *Mov. Disord.* 31, 915–923 (2016).

15. Parkinson Study Group STEADY-PD III Investigators. Isradipine Versus Placebo in Early Parkinson Disease: A Randomized Trial. *Ann Intern Med* 172, 591–598 (2020).
16. Martínez Carrasco, A. *et al.* Genome-wide Analysis of Motor Progression in Parkinson Disease. *Neurol Genet* 9, e200092 (2023).
17. Iwaki, H. *et al.* Genomewide association study of Parkinson's disease clinical biomarkers in 12 longitudinal patients' cohorts. *Mov Disord* 34, 1839–1850 (2019).
18. Hoehn, M. M. & Yahr, M. D. Parkinsonism: onset, progression and mortality. *Neurology* 17, 427–442 (1967).
19. 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. *Mov Disord* 23, 2129–2170 (2008).
20. Nasreddine, Z. S. *et al.* The Montreal Cognitive Assessment, MoCA: a brief screening tool for mild cognitive impairment. *J. Am. Geriatr. Soc.* 53, 695–699 (2005).
21. Jost, S. T. *et al.* Levodopa Dose Equivalency in Parkinson's Disease: Updated Systematic Review and Proposals. *Mov Disord* 38, 1236–1252 (2023).
22. Siderowf, A. *et al.* Assessment of heterogeneity among participants in the Parkinson's Progression Markers Initiative cohort using α-synuclein seed amplification: a cross-sectional study. *Lancet Neurol.* 22, 407–417 (2023).
23. Endo, M. *et al.* Data-Driven Discovery of Movement-Linked Heterogeneity in Neurodegenerative Diseases. *Nat Mach Intell* 6, 1034–1045 (2024).
24. Deng, X. *et al.* Disease Progression of Data-Driven Subtypes of Parkinson's Disease: 5-Year Longitudinal Study from the Early Parkinson's Disease Longitudinal Singapore (PALS) Cohort. *J Parkinsons Dis* 14, 1051–1059 (2024).
25. Inguanzo, A. *et al.* MRI subtypes in Parkinson's disease across diverse populations and clustering approaches. *NPJ Parkinsons Dis* 10, 159 (2024).
26. Lawton, M. *et al.* Developing and validating Parkinson's disease subtypes and their motor and cognitive progression. *J. Neurol. Neurosurg. Psychiatry* 89, 1279–1287 (2018).
27. Su, C. *et al.* Identification of Parkinson's disease PACE subtypes and repurposing treatments through integrative analyses of multimodal data. *NPJ Digit Med* 7, 184 (2024).
28. Cilia, R. *et al.* Natural history of motor symptoms in Parkinson's disease and the long-duration response to levodopa. *Brain* 143, 2490–2501 (2020).
29. Ligaard, J., Sannæs, J. & Pihlstrøm, L. Deep brain stimulation and genetic variability in Parkinson's disease: a review of the literature. *NPJ Parkinsons Dis* 5, 18 (2019).
30. de Oliveira, L. M. *et al.* Deep Brain Stimulation in Patients With Mutations in Parkinson's Disease-Related Genes: A Systematic Review. *Mov Disord Clin Pract* 6, 359–368 (2019).
31. Artusi, C. A. *et al.* Association of Subthalamic Deep Brain Stimulation With Motor, Functional, and Pharmacologic Outcomes in Patients With Monogenic Parkinson Disease: A Systematic Review and

Meta-analysis. *JAMA Netw Open* 2, e187800 (2019).

32. Anis, S. *et al.* Are LRRK2 p.G2019S or GBA1 variants associated with long-term outcomes of deep brain stimulation for Parkinson's disease? *Parkinsonism Relat Disord* 124, 106008 (2024).
33. Yoon, E. *et al.* Association of polygenic risk score with response to deep brain stimulation in Parkinson's disease. *BMC Neurol* 23, 143 (2023).
34. Espay, A. J., Brundin, P. & Lang, A. E. Precision medicine for disease modification in Parkinson disease. *Nat Rev Neurol* 13, 119–126 (2017).
35. Pagano, G. *et al.* Trial of Prasinezumab in Early-Stage Parkinson's Disease. *N Engl J Med* 387, 421–432 (2022).
36. Lang, A. E. *et al.* Trial of Cinpanemab in Early Parkinson's Disease. *N Engl J Med* 387, 408–420 (2022).
37. Brockmann, K. *et al.* CSF α-synuclein seed amplification kinetic profiles are associated with cognitive decline in Parkinson's disease. *NPJ Parkinsons Dis* 10, 24 (2024).
38. Schumacher, J. G. *et al.* Baseline α-synuclein seeding activity and disease progression in sporadic and genetic Parkinson's disease in the PPMI cohort. *EBioMedicine* 119, 105866 (2025).
39. Teixeira-Dos-Santos, D., Tan, A.-H., Okubadejo, N., Kalia, L. V. & Schuh, A. F. S. Bridging global diversity gaps in Parkinson disease research. *Nat Rev Neurol* 22, 239–250 (2026).

Tables

Tables are available in the Supplementary Files section.

Figures

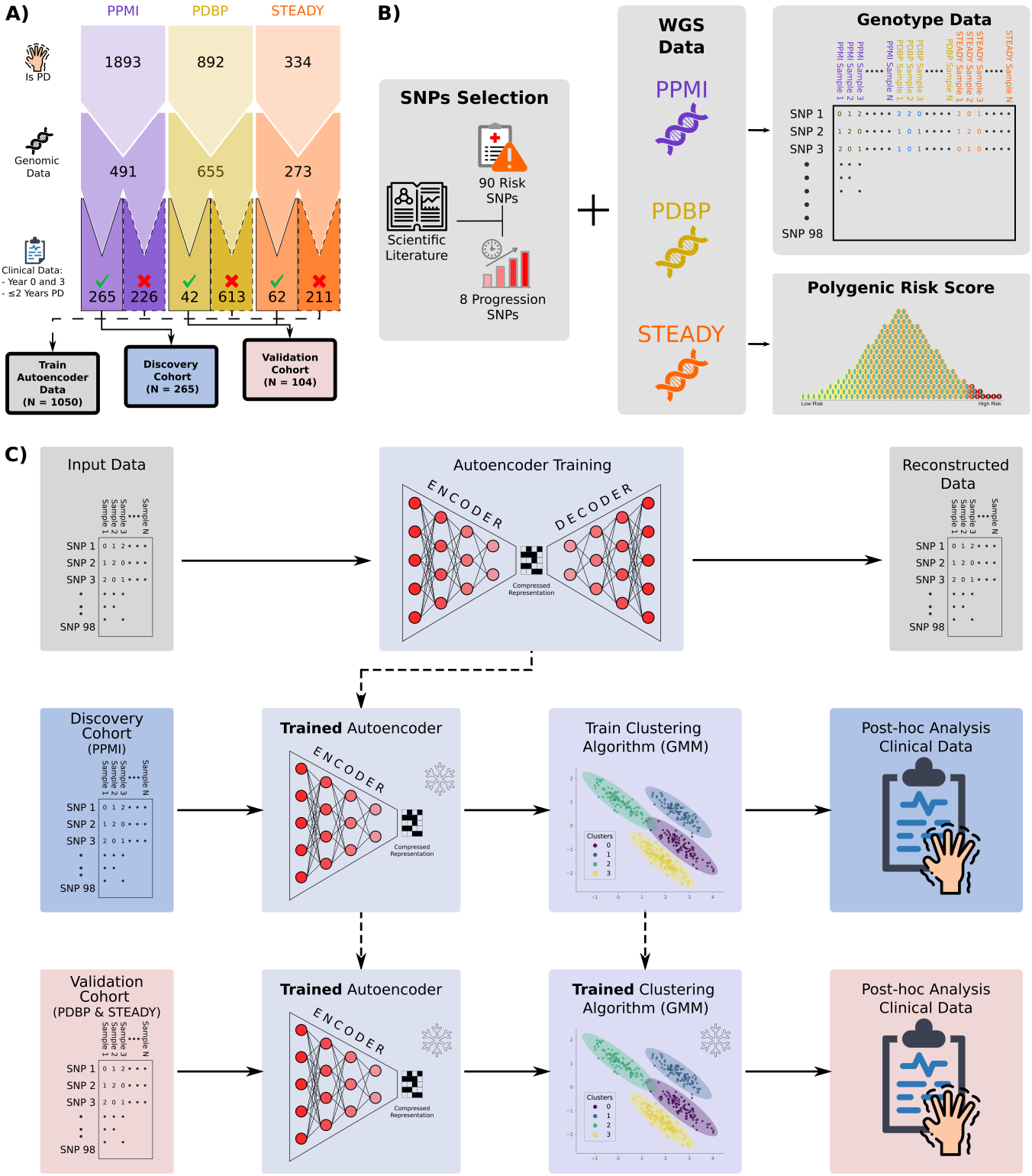


Figure 1

Legend not included with this version.

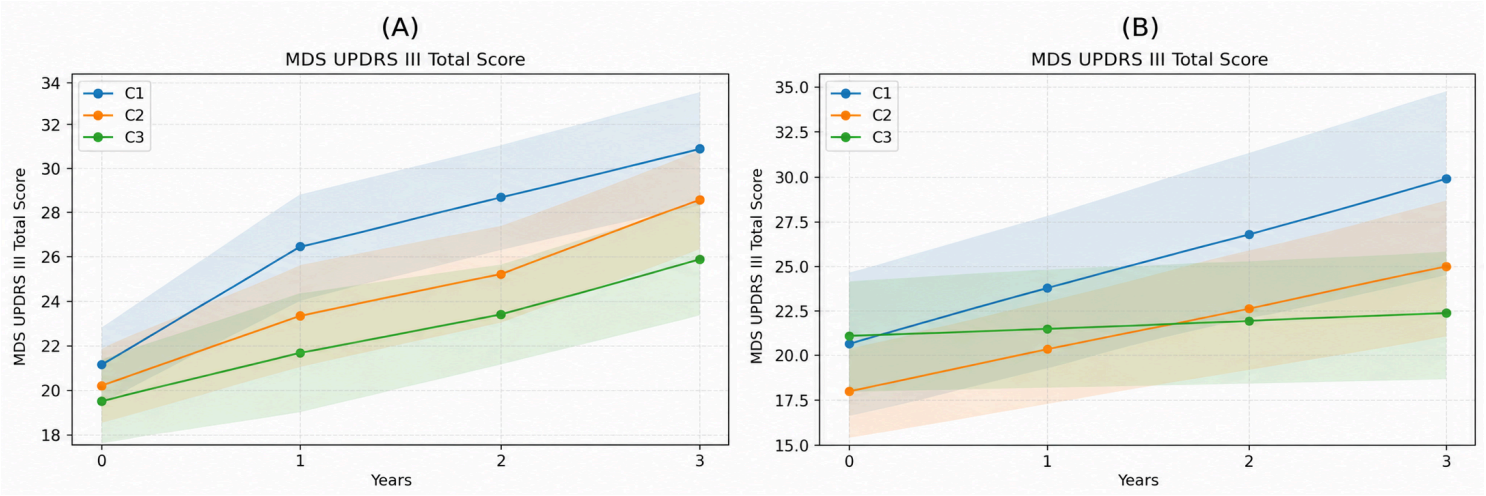
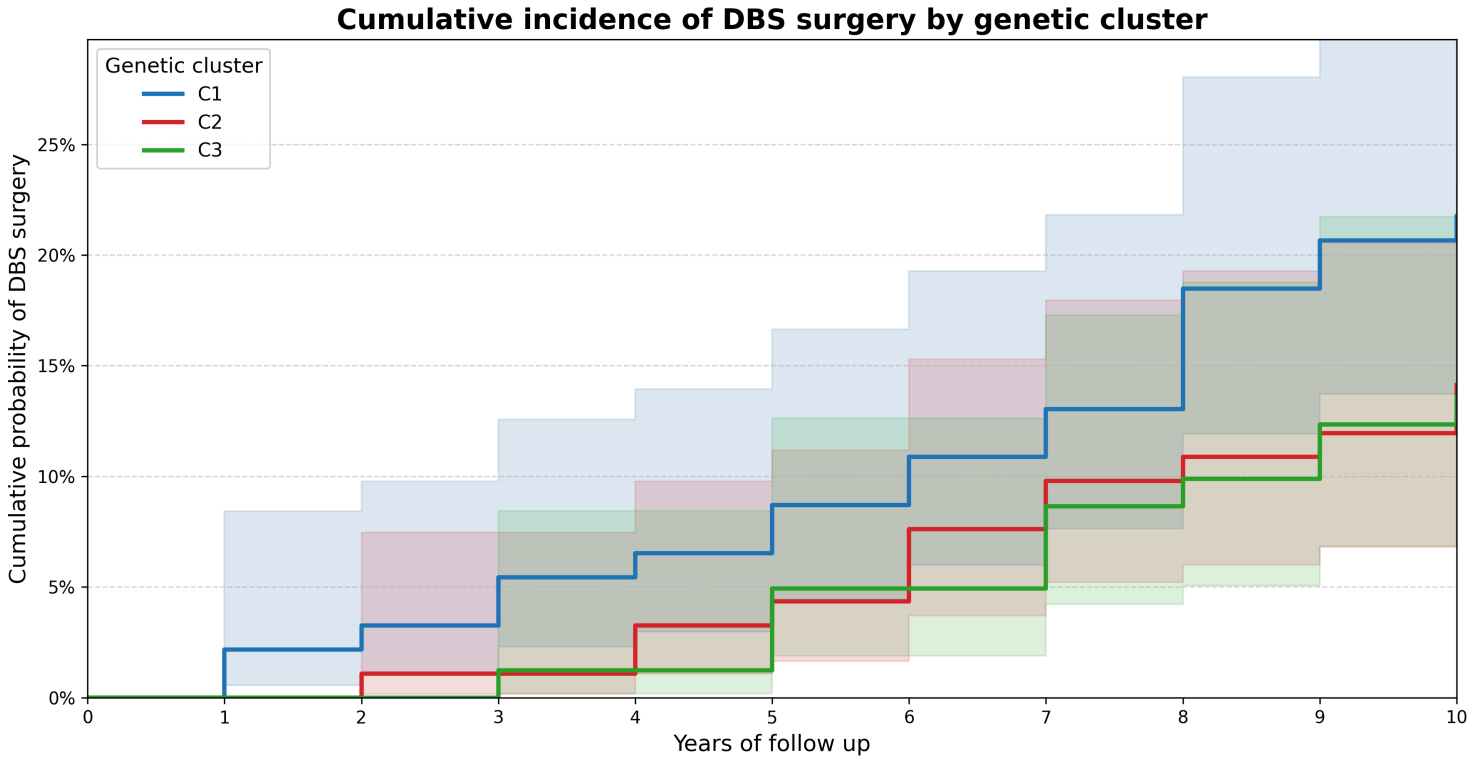


Figure 2

Legend not included with this version.



Cumulative DBS surgeries per year, n

Cluster	0	1	2	3	4	5	6	7	8	9	10
C1	0	2	3	5	6	8	10	12	17	19	20
C2	0	0	1	1	3	4	7	9	10	11	13
C3	0	0	0	1	1	4	4	7	8	10	11

Figure 3

Legend not included with this version.

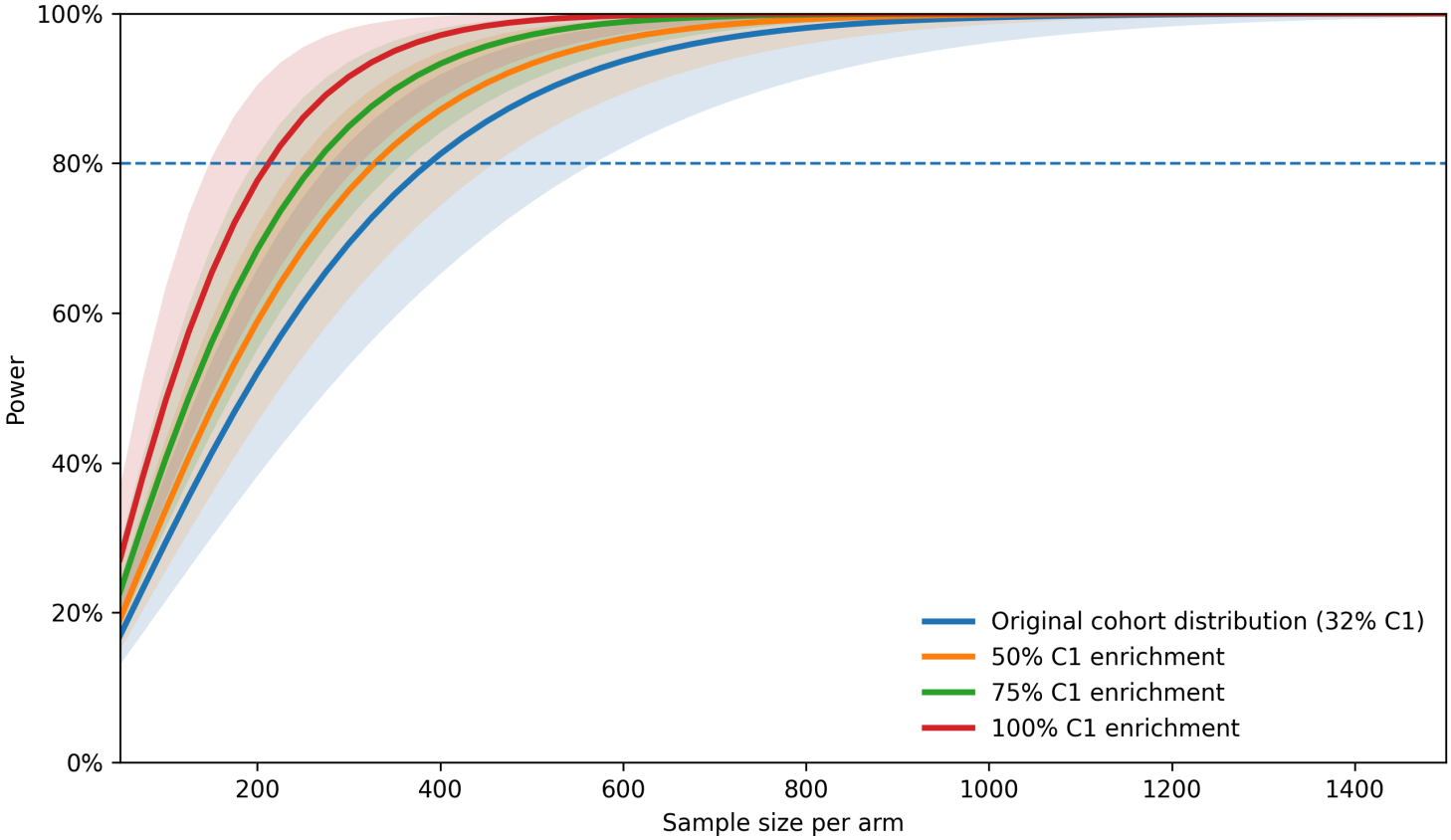


Figure 4

Legend not included with this version.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- Table1.xlsx
- Table2.xlsx
- Table3.xlsx
- SupplementaryMethods.docx
- SupplementaryTables.xlsx