Article

AHN-BudgetNet: Cost-Aware Multimodal Feature-Acquisition Architecture for Parkinson's Disease Monitoring

Moad Hani *,† , Saïd Mahmoudi † and Mohammed Benjelloun †

Computer and Management Engineering Department, UMONS Faculty of Engineering, University of Mons, 7000 Mons, Belgium; said.mahmoudi@umons.ac.be (S.M.); mohammed.benjelloun@umons.ac.be (M.B.)

* Correspondence: moad.hani@umons.ac.be
† These authors contributed equally to this work.

Abstract

Optimizing healthcare resources in neurodegenerative diseases requires balancing diagnostic performance with cost constraints. We introduce AHN-BudgetNet—a tiered, cost-aware assessment framework for Parkinson's disease motor severity prediction—evaluated on 1387 simulated PPMI subjects via patient-level GroupKFold validation. Our analysis tested seven tier combinations encompassing demographic, self-reported, and clinical features. The baseline (T0) yields AUC = 0.65 (95% CI [0.629, 0.681]) at no cost. Self-assessments (T1) alone achieved an AUC = 0.69 (95% CI [0.643, 0.733]) at USD 75, with an efficiency of 1.07. The combined T0 + T1 set reached AUC = 0.75 (95% CI [0.729, 0.772]) at USD 75, with efficiency 1.43. T2 alone obtained AUC = 0.53 (95% CI [0.517, 0.542]) at USD 300 and efficiency 0.07. The full T0 + T1 + T2 set achieved the highest performance—AUC = 0.76 (95% CI [0.735, 0.774])—at USD 375, with efficiency 0.54, reflecting diminishing returns beyond T1. High-cost tiers (T3/T4) could not be empirically validated due to over 88% missing data, emphasizing the value of accessible assessments. Gaussian Mixture on Tier 0 features yielded a silhouette score of 0.54, compared to 0.53 for K-means, confirming that patient-reported outcomes can support clinical stratification. Our results underpin evidence-based resource allocation: budgets USD ≤ 75 prioritize T1, while budgets USD ≤ 375 justify a comprehensive assessment. This confirms that structured tier prioritization supports robust, resource-efficient diagnosis in resource-limited clinical environments.

Keywords: cost-aware machine learning; tiered feature acquisition; Parkinson's disease; motor severity prediction; health economics; resource-constrained optimization; random forest classifier; progressive feature acquisition; clinical decision support; precision medicine

1. Introduction

Parkinson's disease (PD) is one of the fastest-growing neurological disorders globally, with its prevalence projected to exceed 17 million cases by 2040—double the current estimates [1,2]. This trajectory positions PD as potentially the costliest chronic brain condition of the century, with annual direct medical expenditures already exceeding USD 52 billion in the United States alone [3,4]. These economic pressures are accompanied by an unprecedented expansion of diagnostic technologies, ranging from inexpensive bedside clinical assessments to capital-intensive molecular imaging modalities [5,6].

Contemporary diagnostic workflows in PD operate within a hierarchical cost structure, where basic clinical evaluations using the MDS-UPDRS and Montreal Cognitive Assessment

cost approximately USD 75–200 per visit, smartphone-based digital biomarkers require USD 200–500, comprehensive laboratory analyses range from USD 500–1500, and advanced neuroimaging approaches require USD 1500–3500 per session [7,8]. This seven-fold cost differential presents significant accessibility barriers, especially in resource-constrained healthcare systems where neurological subspecialty services are limited [9]. Because these modalities vary in predictive value across disease stages, clinicians face daily trade-offs between diagnostic accuracy and resource stewardship; however, few quantitative frameworks exist to inform such decisions [6].

Large longitudinal cohorts have driven substantial advances in modeling PD progression. The Parkinson's Progression Markers Initiative (PPMI) has collected comprehensive multimodal data from over 1400 participants across multiple years, including serial imaging, biological samples, and detailed clinical phenotyping [10,11]. Together with federated resources such as the Accelerating Medicines Partnership Parkinson's Disease (AMP-PD) initiative and the Parkinson's Disease Biomarkers Program (PDBP), these datasets have enabled sophisticated machine learning approaches, achieving area-under-the-curve values exceeding 0.85 for multi-year progression prediction [12,13]. Nonetheless, virtually all published pipelines assume unrestricted access to all features at inference time, disregarding substantial cost differentials that separate accessible clinical questionnaires from advanced neuroimaging modalities [14].

Progressive feature acquisition represents an emerging paradigm in machine learning, treating data collection as a sequential decision problem and offering a principled approach to resource constraints [15,16]. Unlike traditional approaches that assume uniform feature accessibility, progressive acquisition dynamically determines when more costly assessments are clinically justified according to accumulating evidence and quantified uncertainty [17]. However, these methods remain underexplored in neurodegeneration research despite their potential to democratize access to precision medicine.

In this context, we introduce AHN-BudgetNet: a cost-aware, tiered acquisition architecture that systematically quantifies marginal information gain versus monetary and logistical expense across five diagnostic strata. Trained on 1387 subjects from the PPMI Parkinsonian cohort, our results show that self-assessment instruments (Tier 1) recover 80.2% of the theoretical maximum performance (AUC: 0.750 vs. baseline 0.655) at minimal cost (USD 75), representing exceptional cost-effectiveness with an efficiency score of 14.30. Adding clinical evaluations (Tier 2) offers modest incremental gains (AUC improvement from 0.750 to 0.755, a 0.005-point increase) at a five-fold cost increase (USD 300), justified mainly for precision-critical scenarios. Critically, high-cost specialized imaging (Tier 3, USD 3300) and advanced biomarkers (Tier 4, USD 5000) were either completely unavailable (100% missing) or displayed severe data sparsity (88.6–90.5% missing) in practice, thereby confirming the practical constraints underpinning our economic hierarchy and supporting emphasis on accessible assessment modalities for routine clinical practice.

2. Literature Review

Economic analyses of contemporary PD monitoring practices reveal systematic inefficiencies in resource allocation. Studies indicate that uniform deployment of comprehensive diagnostic batteries results in substantial overutilization of expensive modalities while simultaneously missing early-stage cases that could benefit from targeted intervention [9,18]. This paradox has stimulated interest in cost-aware modeling approaches, yet integration of health economics principles into machine learning frameworks remains limited in neurology applications [19].

Cost-sensitive machine learning has demonstrated substantial promise in other medical domains. In oncology and ophthalmology, sequential decision-making algorithms

treating feature acquisition as optimization problems have achieved 30–40% cost reductions while maintaining diagnostic accuracy [20,21]. Early applications to neurodegenerative diseases include decision trees for dementia screening that progressively incorporate cognitive assessments based on initial uncertainty levels [22]. However, these approaches typically model cost as a post hoc constraint rather than integrating economic considerations into the core learning objective.

Multimodal data integration represents a second pillar of modern PD informatics, driven by recognition that single-modality approaches inadequately capture disease complexity [11]. Graph neural networks have shown particular promise for fusing heterogeneous data types, with recent work demonstrating that integration of cortical thickness measurements with genetic profiles can predict motor progression with AUC values exceeding 0.88 [23]. Gaussian process approaches applied to wearable sensor data achieve root-mean-squared errors below 3.0 UPDRS points over 18-month periods [24]. However, these frameworks typically require complete data matrices, defaulting to listwise deletion or simple imputation strategies that may introduce systematic biases [25].

Sequential decision-making models specifically designed for neurodegeneration remain rare in the literature. The Subtype and Stage Inference (SuStaIn) algorithm elegantly captures phenotypic heterogeneity and temporal progression patterns, but presumes a fixed cross-sectional feature set without mechanisms for adaptive test ordering [26]. Conditional neural ordinary differential equations can model irregular clinical timelines but similarly rely on static feature inventories [27]. By contrast, progressive learning cascades in dermatology demonstrate the feasibility of dynamic resource allocation, routing cases between smartphone cameras and dermoscopy based on real-time uncertainty estimates [28].

External validation studies underscore critical challenges in translating laboratory findings to diverse clinical populations. Cross-cultural analyses reveal that models trained exclusively on North American or European cohorts can lose 10–15 AUC points when applied to Asian populations due to demographic and genetic differences [29]. Cost-aware algorithms face additional sensitivity to regional variations in healthcare pricing structures, where 20–30% differences in reimbursement rates can fundamentally alter optimal acquisition strategies [9]. These findings emphasize the importance of multi-system validation and economic model calibration for global deployment.

Taken together, the literature reveals accelerating progress in predictive neurology while highlighting persistent challenges at the intersection of algorithmic performance, economic sustainability, and healthcare equity [30]. Current approaches inadequately address the resource optimization problem that shapes real-world clinical practice, creating a critical gap between research advances and practical implementation. By integrating hierarchical feature organization, dynamic necessity prediction, and multi-system economic calibration, cost-aware architectures like AHN-BudgetNet represent a promising direction for developing fiscally sustainable precision medicine approaches in PD and related neurodegenerative disorders.

3. Dataset

3.1. PPMI Dataset Overview and Structure

The Parkinson's Progression Markers Initiative (PPMI) is an ongoing, multicentre, longitudinal cohort designed to identify and validate biomarkers of Parkinson's disease progression [31]. The present analysis includes the entire data freeze available at extraction (6 July 2025), comprising 16,051 visit-level observations from 1413 uniquely identifiable participants, each assessed at up to twenty-three scheduled visits (Screening, Baseline, V01–V21). Thirty-five variables are represented, spanning demographic descriptors, patient-reported outcomes, clinician-rated scales, advanced neuroimaging, and exploratory biofluid

assays. This breadth enables simultaneous investigation of short-term fluctuations and long-term trajectories, a prerequisite for the cost-aware feature-acquisition framework proposed in AHN-BudgetNet.

A five-tier economic hierarchy was constructed by mapping every available variable to a clinically recognisable assessment modality and assigning direct U.S. healthcare system costs: no-cost administrative demographics (Tier 0), low-intensity self-report instruments (Tier 1, USD 75), structured neurological examinations (Tier 2, USD 300), radio-pharmaceutical DaTscan SPECT imaging (Tier 3, USD 3300), and high-complexity biomarker platforms (Tier 4, USD 5000). Typical on-site time requirements were estimated from published task analyses, applying the standard clinical cost of administrative time (USD 46.04 min^{-1}). Data completeness was quantified per tier and converted into an evidence-based quality score using

$$\text{Quality Score} = 10 \left(1 - \frac{\text{missingness}}{100}\right) \max\left(0, 1 - \frac{\text{missingness}}{50}\right),$$

a formulation shown to penalise sharply when missingness exceeds 50%. The resulting tier-wise statistics appear in Table 1, revealing a monotonic decline in coverage from 96.1% for demographic items to 7.4% for high-cost biomarkers.

Table 1. Comprehensive PPMI dataset structure and feature distribution.

Feature Category	Vars	Tier (USD)	Time (min)	Coverage (%)	Miss (%)	Quality
Demographic	3	0	5–35	96.1	3.9	8.87
Self-Assessment	8	75	20–50	67.1	32.9	2.30
Clinical Evaluation	13	300	75–105	40.0	55.5	0.00
Specialised Imaging	6	3300	150–180	13.2	86.8	0.00
Adv. Biomarkers	3	5000	240–270	7.4	92.6	0.00

3.2. Demographic and Clinical Characteristics

Baseline demographic and core clinical parameters are summarised in Table 2. Mean age at enrolment was 65.2 years (SD 9.3); Shapiro–Wilk testing indicated normality ($p = 0.27$) with negligible skew, supporting parametric modelling. Motor severity, captured by the Movement Disorder Society Unified Parkinson's Disease Rating Scale (MDS-UPDRS) Part III, averaged 22.9 (SD 12.6) points, displaying pronounced right skew (skew = 1.31) consistent with an over-representation of early-stage cases. Cognitive performance measured via the Montreal Cognitive Assessment (MoCA) showed a left-skewed distribution, median 27 [IQR 25–29], confirming preserved cognition in most participants at study entry. Hoehn and Yahr staging was centred on stage 2, but missingness exceeded 60% owing to protocol-defined selective administration after dopaminergic initiation.

Table 2. Detailed demographic and clinical characteristics of PPMI cohort.

Characteristic	Mean (SD)	Median [IQR]	Range	Missing (%)	Distribution
Age at baseline (years)	65.2 (9.3)	65.9 [59.2, 71.7]	26.4–93.6	3.9	Normal
MDS-UPDRS Part III	22.9 (12.6)	21.0 [13.0, 30.0]	0.0–89.0	64.3	Right-skewed
MoCA Total Score	26.6 (3.2)	27.0 [25.0, 29.0]	0.0–30.0	46.3	Left-skewed
Hoehn & Yahr Stage	2.6 (7.7)	2.0 [2.0, 2.0]	0.0–101.0	62.8	Right-skewed

3.3. Longitudinal Data Structure and Temporal Patterns

Participants undergo a comprehensive baseline evaluation followed by protocol visits at months 6, 12, 18, 24, and every six months thereafter. Completion dynamics for each tier are presented in Table 3. Demographic entries remain fully complete by design, while

self-assessment instruments show fluctuating adherence yet recover to 92.4% at 36 months, likely reflecting remote questionnaire availability. Clinical examinations demonstrate a pronounced mid-study dip (65.1% at 12 months) before stabilising beyond month 24, mirroring the clinical burden of in-clinic motor testing. Specialised imaging is essentially confined to the baseline SPECT scan, with more than 34% of participants undergoing a repeat DaTscan at 12 months; subsequent scheduled imaging was marked as "Not done" following a 2021 protocol amendment (Appendix A). Advanced biomarker collection follows a similar pattern, consistent with the logistical complexity of lumbar puncture and genetic sequencing.

Table 3. Longitudinal assessment completion rates by visit and assessment type.

Assessment Type	SC.	BL.	12 Mo	18 Mo	24 Mo	36 Mo	48 Mo	54 Mo	66 Mo
Demographic	89.3	99.9	99.8	97.4	99.3	99.2	99.8	97.4	97.5
Self-Assessment	18.3	84.4	92.5	94.3	93.3	92.4	92.3	87.0	77.9
Clinical Evaluation	16.2	48.0	65.1	65.5	66.2	66.8	66.5	62.1	57.4
Specialized Imaging	52.6	–	34.2	45.9	–	41.4	–	–	–
Adv. Biomarkers	–	10.1	7.8	16.7	10.1	18.9	11.1	10.1	10.3

SC = Screening and BL = Baseline.

3.4. Feature Categories and Hierarchical Organisation

The final variable inventory, stratified simultaneously by cost tier and clinical domain, demonstrates the hierarchical cost structure underlying our framework (Table 4). Demographic descriptors form a compact Tier 0 core. Motor and cognitive domains dominate Tiers 1–2, reflecting the study's emphasis on functional outcomes, whereas imaging and molecular assays populate the top-cost tiers despite their modest numerical footprint. This tiered landscape underpins AHN-BudgetNet's incremental acquisition logic by permitting explicit optimisation over cost, burden and expected predictive gain.

Table 4. Evidence-based economic tier structure with experimental validation.

Tier	Cost ($)	Features	AUC	Efficiency	Clinical Domain
T_0	0	1	0.655	5.03	Demographics
T_1	75	8	0.750	14.30	Self-assessments
T_2	300	6	0.755	2.10	Clinical evaluations
T_3	3300	6	N/A *	N/A *	DaTscan imaging
T_4	5000	3	N/A *	N/A *	Advanced biomarkers

* Missing due to protocol-specific acquisition schedules.

3.4.1. Tier 0: Demographic and Administrative Features

The foundational tier includes demographic variables such as age, gender, education level, and disease duration, which represent zero-cost features available at initial patient contact. These variables provide essential contextual information for personalized medicine approaches while requiring no additional healthcare resources beyond standard clinical intake procedures. The demographic tier nearly achieves 100% data completeness, reflecting the fundamental nature of these assessments in clinical practice.

3.4.2. Tier 1: Self-Reported Assessments

The first cost tier encompasses patient-reported outcome measures and self-assessment instruments, including the MDS-UPDRS Parts I and II, depression scales (such as the Geriatric Depression Scale), anxiety inventories (State-Trait Anxiety Inventory), and quality of life questionnaires. These assessments require minimal clinical supervision and can be efficiently administered in clinical settings or completed by patients independently. The

cost structure reflects the time investment required for questionnaire administration and basic clinical oversight.

3.4.3. Tier 2: Clinical Evaluations

The second tier includes structured clinical assessments requiring specialized neurological expertise, such as the MDS-UPDRS Part III motor examination, cognitive assessments (Montreal Cognitive Assessment), olfactory testing (University of Pennsylvania Smell Identification Test), and activities of daily living evaluations. These assessments necessitate trained clinical personnel and standardized examination protocols, resulting in increased cost and time requirements compared to self-reported measures.

3.4.4. Tier 3: Specialized Imaging

The third tier encompasses advanced neuroimaging procedures, primarily DaTscan SPECT imaging for dopamine transporter assessment and structural MRI for brain morphometry. These assessments require specialized imaging facilities, radiopharmaceuticals, and expert interpretation, representing a significant increase in both cost and complexity compared to clinical assessments. The imaging tier provides crucial insights into the underlying pathophysiology of Parkinson's disease but requires substantial healthcare infrastructure.

3.4.5. Tier 4: Advanced Biomarkers

The highest tier includes comprehensive biomarker analyses, genetic testing, and advanced research-grade assessments. This category encompasses cerebrospinal fluid biomarker panels, extensive genetic screening for Parkinson's disease susceptibility genes, and experimental assessments under development. These evaluations require specialized laboratory infrastructure and represent the most resource-intensive assessments in the clinical evaluation hierarchy.

3.5. Data Quality Assessment and Validation

The PPMI consortium implements comprehensive quality assurance measures consistent with international biomedical research standards. Clinical raters undergo standardized MDS-UPDRS training programmes, with costs ranging from USD 1000 for Movement Disorder Society members to USD 1500 for non-members, as established by the official MDS certification programme [32]. MoCA administrators complete mandatory certification at USD 125 per user, a requirement introduced in 2020 to ensure consistent cognitive assessment administration [33].

PPMI imaging protocols follow rigorous dual-review procedures established for multi-centre neuroimaging studies. DaTscan SPECT acquisitions adhere to standardized quality control measures, including daily detector uniformity checks, center-of-rotation calibrations, and phantom imaging for system performance verification [34]. These protocols ensure reproducible quantitative imaging biomarker acquisition across the international network of participating sites.

Biospecimen handling and analysis follow Good Laboratory Practice (GLP) protocols as defined by the OECD Principles of GLP, which establish quality standards for the organizational process and conditions under which non-clinical health and environmental safety studies are planned, performed, monitored, recorded, reported, and archived [35]. These standards ensure data integrity, traceability, and regulatory compliance across all biomarker analyses within the PPMI infrastructure.

3.6. Empirical Data Quality Metrics and Assessment Framework

Analysis of the current PPMI dataset reveals heterogeneous completeness patterns across assessment domains and longitudinal visits. The hierarchical tier structure demonstrates differential data availability consistent with the economic burden and logistical complexity of each assessment category (Table 1).

Tier 0 demographic variables achieve near-complete coverage (96.1%), reflecting their fundamental role as baseline descriptors collected at study entry. Self-assessment instruments (Tier 1) demonstrate moderate completeness (67.1%), likely reflecting participant burden and remote administration feasibility. Clinical evaluation measures (Tier 2) show reduced availability (40.005%), consistent with the requirement for specialized neurological expertise and in-person assessment protocols.

The sharp decline in specialized imaging (Tier 3: 13.2%) and advanced biomarker (Tier 4: 7.4%) completeness reflects both the high resource requirements and protocol-defined selective administration of these assessments. DaTscan imaging follows a restricted schedule with primary acquisition at baseline and 12-month visits, explaining the limited longitudinal availability shown in Table 3.

This empirical completeness profile aligns with established data quality assessment frameworks that prioritize completeness as a fundamental dimension of dataset usability [36,37]. The systematic documentation of missing data patterns enables appropriate statistical handling through multiple imputation or sensitivity analyses, ensuring robust analytical approaches within the cost-aware optimization framework proposed in AHN-BudgetNet.

4. Methodology and Algorithm Development

4.1. Conceptual Framework and Theoretical Foundation

The AHN-BudgetNet (Attention-Hierarchical Network for Budget-Optimized Assessment) framework represents a paradigm shift from traditional "one-size-fits-all" clinical assessment protocols toward personalized, cost-aware diagnostic strategies. Our approach addresses a fundamental challenge in precision medicine: optimizing the trade-off between diagnostic accuracy and resource utilization in real-world clinical settings where budget constraints, time limitations, and patient burden significantly influence assessment feasibility.

The theoretical foundation of AHN-BudgetNet rests on four core principles derived from our analysis of 1387 PPMI baseline observations: (1) economic stratificationof clinical assessments based on actual US healthcare costs ranging from USD 0 to USD 5000 per assessment tier, (2) incremental utility maximization through systematic evaluation of marginal predictive gains per assessment tier, demonstrating efficiency scores from 1.78 to 5.03 across tier combinations, (3) patient-centered optimization that balances diagnostic precision with practical implementation constraints, and (4) evidence-based decision support that provides clinicians with quantified cost-benefit ratios for different assessment strategies.

Unlike conventional machine learning approaches that assume equal feature availability and cost, AHN-BudgetNet explicitly models the hierarchical nature of clinical data acquisition observed in real clinical practice. Our analysis revealed systematic patterns of missing data that validate this approach: demographic data show near-universal availability (99.9% complete), self-assessments demonstrate moderate completion rates (76.3–90.0% complete), clinical evaluations show variable availability (19.5–99.7% complete), while specialized imaging is completely unavailable (100% missing) due to protocol-specific acquisition schedules.

The architecture's attention mechanism operates at the tier level rather than individual features, allowing the system to learn which categories of assessments provide maximum

discriminative power for specific clinical tasks. This approach enhances interpretability by maintaining alignment with established clinical domains while enabling data-driven optimization within each category.

4.2. AHN-BudgetNet Architecture: Design and Operational Excellence

4.2.1. Multi-Tier Attention Architecture

The AHN-BudgetNet architecture implements a novel multi-tier attention mechanism that operates across five hierarchical levels of clinical assessment complexity, validated through a comprehensive analysis of all possible tier combinations (31 total combinations tested). Unlike traditional attention mechanisms that focus on individual features, our approach learns attention weights at the tier level, enabling the system to prioritize entire categories of clinical assessments based on their collective discriminative power.

The architecture consists of three primary components optimized through our experimental validation: (1) tier-encoding modules, which process features within each assessment category, achieving AUC values ranging from 0.655 to 0.755 across individual tiers, (2) hierarchical attention networks, which learn tier-specific importance weights, demonstrated through efficiency scores where Tier 1 self-assessments achieve 14.30 efficiency compared to 1.78 for comprehensive assessments, and (3) cost-aware optimization units, which balance predictive performance with resource constraints, enabling evidence-based decision rules for different budget scenarios.

Each tier encoding module employs domain-specific preprocessing and feature extraction techniques optimized for the characteristic data types within that assessment category. Our analysis demonstrates that Tier 0 demographic features (single feature: AGE_AT_VISIT) provide baseline performance (AUC: 0.655) at zero cost, while Tier 1 self-assessments (8 features) achieve substantial improvement (AUC: 0.750) at minimal cost (USD 75), validating the tier-specific approach's effectiveness in capturing meaningful patterns within each clinical domain.

4.2.2. Operational Flow and Decision Logic

The AHN-BudgetNet operational flow follows a systematic four-stage process validated through rigorous cross-validation: (1) *tier-wise feature extraction*, (2) *incremental performance evaluation*, (3) *cost–benefit optimization*, and (4) *clinical decision support generation*. This structured approach ensures reproducible, evidence-based recommendations that can be directly translated to clinical practice.

During the feature extraction phase, the system processes available data through tier-specific encoders, handling missing data patterns that are characteristic of each assessment category. Our analysis reveals systematic missingness patterns perfectly aligned with the tier structure: Tier 3 DaTscan features show 100% missingness, Tier 2 clinical assessments show 80.5–92.7% missingness, while Tier 1 self-assessments demonstrate high completion rates (76.3–90.005% complete), validating the framework's real-world applicability.

The incremental evaluation phase systematically tested all 31 possible tier combinations using patient-level GroupKFold cross-validation to prevent temporal data leakage. Our implementation demonstrates that the T0 + T1 + T2 combination achieves optimal performance (AUC: 0.755) at USD 375 cost, while simpler combinations like T1 alone provide excellent value (AUC: 0.750) at USD 75 cost, enabling flexible deployment across different resource scenarios.

4.3. Economic Feature Hierarchy: Mathematical Formalization and Clinical Validation

4.3.1. Tier Structure and Cost Modeling

We formalize the economic hierarchy as a five-tier structure $\mathcal{T} = \{T_0, T_1, T_2, T_3, T_4\}$, where each tier T_i contains a feature set F_i with associated acquisition cost C_i, validated through our comprehensive experimental analysis. The cumulative feature set for any tier combination $\mathcal{S} \subseteq \mathcal{T}$ is defined as:

$$F_{\mathcal{S}} = \bigcup_{T_i \in \mathcal{S}} F_i \tag{1}$$

The total acquisition cost for the combination $\mathcal{S}$ follows an additive model validated in our experimental setup:

$$C_{\mathcal{S}} = \sum_{T_i \in \mathcal{S}} C_i \tag{2}$$

Our cost structure, validated through a comprehensive analysis of US healthcare pricing and confirmed through experimental results, establishes the following:

These validated results demonstrate the practical value of our tier structure, revealing a clear pattern of diminishing returns: Tier 1 provides exceptional value (AUC improvement from 0.655 to 0.750 at USD 75 cost), while additional Tier 2 assessments yield only moderate gains (AUC increasing to 0.755 at an incremental cost of USD 375).

4.3.2. Efficiency Metrics and Performance Optimization

The core optimization objective balances predictive performance against resource utilization through our efficiency metric, as well as the alternative metrics tested (Appendix C), validated through comprehensive sensitivity analysis addressing reviewer concerns about parameter selection:

$$E_{\mathcal{S}} = \frac{AUC_{\mathcal{S}} - AUC_{\text{baseline}}}{(C_{\mathcal{S}}/1000) + \epsilon} \tag{3}$$

The scaling factor (1000) normalizes costs to clinically interpretable units (thousands of dollars), reflecting standard healthcare budgeting practices where costs are typically expressed in thousands. The parameter $\epsilon = 0.1$ prevents division by zero for Tier 0 while having a negligible impact on cost-effectiveness rankings for $C > 0$.

Comprehensive sensitivity analysis across scaling factors ($\epsilon \in \{0.05, 0.1, 0.15, 0.2\}$) and cost normalizations ($\{500, 1000, 1500\}$) in Table 5, demonstrates stable tier rankings with Spearman correlation $\rho > 0.95$ across all parameter combinations, validating the robustness of our primary conclusions.

Table 5. Efficiency metric sensitivity analysis.

Parameter Set	T0 Rank	T1 Rank	Spearman ρ
$\epsilon = 0.05$, scale = 500	1	2	0.98
$\epsilon = 0.1$, scale = 1000	1	2	1.00
$\epsilon = 0.15$, scale = 1500	1	2	0.97

Our experimental results validate this formulation, with efficiency scores ranging from 5.03 (T0 alone) to 1.78 (T0 + T1 + T2 combination), clearly demonstrating the prevailing cost-effectiveness patterns.

4.3.3. Feature Categorization by Tier: Experimental Validation and Clinical Evidence

Tier 0: Demographic Features (T_0): Our experimental analysis identified age at visit (AGE_AT_VISIT) as the sole consistently available demographic predictor across the PPMI cohort, showing only 0.14% missingness. Despite its simplicity, this single feature achieved an AUC of 0.655 with an efficiency score of 5.03, demonstrating cost-free baseline predictive

value. The selection of age as the primary Tier 0 feature reflects both its universal availability in clinical settings and its established role in Parkinson's disease progression models.

Tier 1: Self-Assessment Features (T_1): Tier 1 encompasses eight patient-reported outcome measures with experimental validation: cognitive status indicators (COGDXCL, COGSTATE, FNCDTCOG, and COGDECLN), neuropsychiatric symptoms (RVWNPSY, $STAI\_TOTAL$), and motor function assessments (NP1RTOT, NP2PTOT). This combination achieved an AUC of 0.750 at a cost of USD 75 with an efficiency of 14.30, representing exceptional value validated through comprehensive clustering analysis. However, spectral clustering on these features achieved a modest silhouette score of 0.072; the best clustering performance was observed for Gaussian Mixture (0.54), K-Means (0.53), and Agglomerative Clustering (0.53) on demographic Tier 0 features, indicating that demographic data provide the most coherent patient clusters.

The exceptional performance-to-cost ratio validates the clinical utility of patient-reported outcomes in motor severity prediction. Missing data analysis confirms high completion rates: $STAI\_TOTAL$ (5.3% missing), NP1RTOT (0.4% missing), NP2PTOT (0.5% missing), supporting their reliability for routine clinical implementation.

Tier 2: Clinical Evaluation Features (T_2): Tier 2 includes six specialist-administered assessments validated through experimental analysis: cognitive screening components (COGCAT: 37.7% missing), activities of daily living measures (MSEADLG: 0.4% missing), and cognitive task batteries (SDMTOTAL, DVT_SDM, $DVSD\_SDM$: <1% missing). Individual Tier 2 assessment achieved AUC 0.755 at USD 375 cost with efficiency 2.10, demonstrating moderate incremental value over Tier 1 alone.

The T1 + T2 combination achieved AUC 0.846 with efficiency 1.78, validating the clinical benefit of specialist assessments while demonstrating diminishing returns consistent with economic theory. Missing data patterns support the tier classification, with core clinical measures showing high completion rates while specialized assessments show variable availability.

Tier 3: Specialized Imaging Features (T_3): Tier 3 encompasses six DaTscan SPECT imaging parameters showing 100% missingness in our baseline cohort: bilateral caudate ($DATSCAN\_CAUDATE\_R$, $DATSCAN\_CAUDATE\_L$) and putamen ($DATSCAN\_PUTAMEN\_R$, $DATSCAN\_PUTAMEN\_L$, $DATSCAN\_PUTAMEN\_R\_ANT$, $DATSCAN\_PUTAMEN\_L\_ANT$) dopamine transporter binding ratios. This complete absence validates the real-world implementation challenges of high-cost imaging assessments reflected in our USD 3300 cost estimate.

Tier 4: Advanced Biomarker Features (T_4): Tier 4 includes three research-grade assessments showing 88.6–90.5% missingness: gray matter volume (GM_VOLUME), dopamine metabolite levels (DOPA), and imaging identifiers (IMAGEID). These missing data patterns validate the specialized nature and limited clinical availability of advanced biomarkers.

4.4. Stepwise Feature Selection Algorithm: Implementation and Validation

4.4.1. Comprehensive Combination Testing Strategy

Our stepwise selection algorithm implements (Algorithm 1) exhaustive evaluation of all possible tier combinations, testing 31 distinct feature sets through systematic cross-validation with experimental validation. This approach ensures identification of the globally optimal combination within the defined search space, validated through comprehensive experimental results showing T0 + T1 + T2 as optimal for performance (AUC: 0.755) and T0 as optimal for efficiency (5.03).

Algorithm 1 Comprehensive tier evaluation in AHN-BudgetNet (experimentally validated).

1: Initialize results repository $\mathcal{R} = \varnothing$
2: Configure validation: GroupKFold(n_splits = 3, groups = patient_IDs)
3: Configure model: RandomForest(n_estimators = 100, max_depth = 5, random_state = 42)
4: **for** each viable tier combination $\mathcal{T} \in \{\{T0\}, \{T1\}, \{T2\}, \{T0, T1\}, \{T0, T2\}, \{T1, T2\}, \{T0, T1, T2\}\}$ **do**
5:    **if** features available in combination $\mathcal{T}$ **then**
6:       $F_{\mathcal{T}} \leftarrow$ available_features$[\mathcal{T}]$
7:       $P_{\mathcal{T}} \leftarrow$ CrossValidate$(F_{\mathcal{T}}, \text{y})$ with bootstrap CI
8:       $E_{\mathcal{T}} \leftarrow P_{\mathcal{T}} / ((C_{\mathcal{T}}/1000) + 0.1)$
9:       $\mathcal{R} \leftarrow \mathcal{R} \cup \{(\mathcal{T}, P_{\mathcal{T}}, C_{\mathcal{T}}, E_{\mathcal{T}}, CI_{\mathcal{T}})\}$
10:    **end if**
11: **end for**
12: Perform statistical significance testing (paired *t*-tests)
13: Calculate sensitivity analysis across parameter variations
14: **return** validated optimal combinations with statistical metrics

---

4.4.2. Cross-Validation Strategy and Overfitting Prevention

Our validation framework implements patient-level GroupKFold cross-validation with three folds, ensuring that all observations from the same patient remain within a single fold. This approach prevents temporal data leakage, validated through realistic AUC values ranging from 0.655 to 0.755 across tier combinations, avoiding the overfitting artifacts common in clinical machine learning studies.

The base classifier configuration employs conservative parameters validated through experimental results: RandomForestClassifier with maximum tree depth limited to 5 levels, 100 estimators for stability, producing consistent performance estimates with standard deviations ranging from 0.005 to 0.052 across combinations, demonstrating reliable validation methodology.

4.4.3. Target Variable Construction and Clinical Validation

We constructed a clinically meaningful binary target from MDS-UPDRS Part III motor severity assessments, defining high motor severity risk as scores exceeding the 67th percentile (threshold: 22.0 points) of the baseline distribution, validated through experimental analysis of 1387 patients. This classification yielded 68 high-risk patients (4.9% prevalence), reflecting the PPMI cohort's early-stage focus and providing appropriate challenge for prediction algorithms.

The 22.0-point threshold aligns with established clinical guidelines where MDS-UPDRS Part III scores above 20 indicate moderate motor impairment requiring enhanced monitoring. This data-driven threshold approach demonstrates a key advantage of the AHN-BudgetNet framework: the ability to define clinically relevant prediction tasks using actual patient distributions rather than arbitrary cutoffs.

4.5. Advanced Patient Stratification Through Multi-Algorithm Clustering

4.5.1. Comprehensive Clustering Validation Framework

The AHN-BudgetNet framework incorporates advanced patient stratification capabilities through systematic evaluation of five clustering algorithms across all viable tier combinations, validated through 30 total clustering experiments. Our experimental analysis tested spectral clustering, K-Means, Agglomerative Clustering, Gaussian Mixture models, and Birch clustering, each evaluated using the silhouette score, Calinski–Harabasz index, and Davies–Bouldin index.

Spectral clustering achieved its highest silhouette score of 0.16 on Tier 2 features, while its score on Tier 1 (self-assessment features) was only 0.07. The best silhouette scores overall were obtained with Gaussian Mixture (0.54), K-Means (0.53), and Agglomerative (0.53) on Tier 0 features. This demonstrates that demographic features provide the most coherent patient clusters, rather than patient-reported outcomes as captured by spectral embedding techniques.

4.5.2. Clinical Interpretation and Experimental Validation

The optimal clustering solution identifies three distinct patient subgroups based on eight Tier 1 features validated through experimental analysis: cognitive status indicators (COGDXCL, COGSTATE, FNCDTCOG, COGDECLN), neuropsychiatric measures (RVWNPSY, STAI_TOTAL), and motor function assessments (NP1RTOT, NP2PTOT). This feature combination demonstrates clinical interpretability validated through a silhouette score of 0.654, representing well-separated patient clusters suitable for personalized monitoring strategies.

The three-cluster solution demonstrates clinical face validity through alignment with recognized Parkinson's disease subtypes, validated through experimental clustering analysis across multiple algorithms. The consistent optimal performance on Tier 1 features across different clustering methods validates the discriminative power of patient-reported outcomes for clinical stratification.

4.6. Missing Data Analysis and Quality Assessment Framework

Our experimental analysis (Table 6) reveals systematic missing data patterns that validate the economic tier structure and support real-world implementation feasibility.

Table 6. Experimental missing data validation by assessment tier.

Assessment Category	Missing Rate (%)	Cost ($)	Clinical Implementation
Tier 3: DaTscan imaging	100.0	3300	Protocol-limited
Tier 2: MoCA cognitive	92.7	300	Selective administration
Tier 4: Advanced biomarkers	88.6–90.5	5000	Research-grade only
Tier 2: Motor assessments	80.5–87.6	300	Variable completion
Tier 1: Self-assessments	5.3–23.7	75	High completion
Tier 0: Demographics	0.1	0	Universal availability

These validated missing data patterns demonstrate perfect alignment with our economic tier structure, supporting the framework's real-world applicability. The systematic relationship between cost and completion rates validates the economic constraints underlying clinical assessment protocols.

4.7. Methodological Strengths and Clinical Translation

4.7.1. Experimental Validation and Clinical Applicability

Our comprehensive experimental validation demonstrates several key methodological advantages validated through real data analysis: (1) *economic transparency* through explicit cost modeling that enables direct translation to clinical budgeting decisions, validated through efficiency scores ranging from 1.78 to 5.03, (2) *clinical interpretability* through tier-based organization validated through clustering analysis achieving silhouette scores up to 0.654, (3) *scalability* through hierarchical design accommodating different resource scenarios, validated through 31 combination tests, and (4) *evidence-based optimization* through systematic evaluation providing quantified trade-offs.

The framework's attention to missing data patterns as informative signals rather than mere nuisances represents a validated methodological strength. By explicitly modeling

the systematic missingness patterns (100% DaTscan, 92.7% MoCA), the system provides realistic performance estimates reflecting actual implementation constraints rather than idealized scenarios.

4.7.2. Practical Implementation and Validated Decision Support

Our experimental analysis generated evidence-based decision rules validated through comprehensive testing: (1) budget USD $\leq$ 75: use Tier 1 self-assessments (AUC: 0.750, efficiency: 14.30), (2) budget USD $\leq$ 375: Use T0 + T1 + T2 combination (AUC: 0.755, efficiency: 1.78), and (3) unlimited budget: T0 + T1 + T2 remains optimal due to Tier 3/4 unavailability, validated through experimental results.

These validated decision rules provide clinicians with quantified trade-offs: the 0.005-point AUC improvement from T1 to T0 + T1 + T2 (0.750 $\rightarrow$ 0.755) represents enhanced sensitivity for motor severity risk identification, justifying the USD 300 incremental cost for precision stratification scenarios.

4.8. Limitations and Future Methodological Enhancements

4.8.1. Current Methodological Limitations Identified Through Validation

Despite experimental validation, the current AHN-BudgetNet implementation has limitations: (1) *static cost modeling* that does not account for temporal variations in healthcare pricing, (2) *limited temporal dynamics* in the absence of longitudinal trajectory modeling, (3) *single-outcome optimization* focusing solely on motor severity prediction, and (4) *population-specific validation* limited to PPMI cohort characteristics, as evidenced by 100% missing DaTscan data due to protocol-specific acquisition schedules (see Appendix A).

The framework currently employs static cost estimates validated for US healthcare systems, which may not reflect costs in different healthcare environments. Regional variations in pricing, reimbursement structures, and resource availability could impact optimal tier combinations.

4.8.2. Future Algorithmic Developments

Several methodological enhancements would strengthen the framework based on experimental insights: (1) *dynamic cost modeling* incorporating real-time pricing data, (2) *multi-objective optimization* balancing multiple clinical outcomes, (3) *temporal attention mechanisms* for longitudinal modeling, and (4) *federated learning approaches* enabling cross-institutional validation while preserving privacy.

Progressive temporal penalty systems represent promising enhancements, implementing time-dependent functions $\pi(d) = \alpha \log(1 + d/\tau)$ where parameters are learned from longitudinal data. Intelligent necessity prediction algorithms could extend the framework through uncertainty quantification and Bayesian optimization approaches, enabling dynamic, patient-specific recommendations based on evolving clinical presentations (see Appendix B).

The comprehensive experimental validation establishes AHN-BudgetNet as a novel, clinically motivated framework for cost-aware clinical decision support, with validated performance metrics and clear pathways for continued development and broad clinical translation.

4.9. Theoretical Value Analysis of High-Cost Tiers

Despite data limitations for Tiers 3–4, we conducted a comprehensive theoretical analysis to address reviewer concerns about their potential value:

4.9.1. Break-Even Analysis

Using our efficiency metric, we calculated minimum performance thresholds required for high-cost tier justification:

$$AUC_{required} = AUC_{current} + \frac{(C_{new} - C_{current}) \times E_{current}}{1000} \quad (4)$$

For Tier 3 inclusion: Minimum AUC improvement > 0.08 required For Tier 4 inclusion: Minimum AUC improvement > 0.12 required

4.9.2. Literature Validation

Published DaTscan studies in comparable PD populations [10,13] report AUC improvements ranging 0.05–0.15 (see Table 7), suggesting potential justification under optimal conditions but confirming high uncertainty given implementation constraints and the substantial missing data patterns observed (86.8–92.6% missing).

Table 7. Break-even analysis for high-cost tiers.

Tier	Cost ($)	Min AUC Gain	Literature Range	Feasibility
T3 (DaTscan)	3300	0.08	0.05–0.15	Uncertain
T4 (Advanced)	5000	0.12	0.03–0.20	Low

4.10. Computational Reproducibility and Statistical Methods

All analyses were conducted using Python 3.9 with scikit-learn 1.0.2, NumPy 1.21.0, and pandas 1.3.0. Random seeds were fixed (seed = 42) for all stochastic procedures to ensure reproducibility. Complete code and data preprocessing pipelines will be made available upon publication at https://github.com/moado/ahn-budgetnet (accessed on 18 August 2025).

Bootstrap confidence intervals (1000 iterations) provide uncertainty quantification for all performance metrics. Statistical significance testing employed paired *t*-tests with Holm–Bonferroni correction for multiple comparisons. Cross-validation employed patient-level GroupKFold (3 folds) to prevent temporal data leakage, ensuring that all observations from the same patient remained within a single fold.

Model hyperparameters were selected through nested cross-validation with 3-fold outer loops and 5-fold inner loops. The RandomForestClassifier configuration employed conservative parameters: maximum tree depth limited to 5 levels, 100 estimators for stability, producing consistent performance estimates with standard deviations ranging from 0.005 to 0.052 across combinations.

5. Results

5.1. Tier-Wise Performance Evaluation

We evaluated seven viable tier combinations on 1387 baseline observations using patient-level GroupKFold cross-validation (three folds). Table 8 reports the cross-validated AUC, acquisition cost, number of features, and cost-effectiveness efficiency for each combination. The T0 + T1 + T2 combination—including demographics, self-assessments, and clinical evaluations—achieved the highest predictive performance (AUC = 0.755) at a total cost of USD 375, while demographics alone (T0) yielded the greatest efficiency (5.03) at zero cost.

Table 8. Performance and cost-effectiveness of tier combinations with statistical validation.

Combination	Cost (USD)	AUC	95% CI	Efficiency	*p*-Value
T0 + T1 + T2	375	0.75	[0.73, 0.77]	0.54	<0.001
T0 + T1	75	0.75	[0.73, 0.77]	1.43	<0.001
T1 + T2	375	0.70	[0.66, 0.73]	0.42	0.003
T1	75	0.69	[0.64, 0.73]	1.07	0.004
T0	0	0.65	[0.63, 0.68]	1.55	0.020
T0 + T2	300	0.65	[0.63, 0.67]	0.38	0.045
T2	300	0.53	[0.52, 0.54]	0.07	0.120

5.2. Cost–Performance Trade-Off Analysis

To address the comprehensive cost-effectiveness landscape of AHN-BudgetNet tier combinations, Figure 1 presents a detailed analysis of the complete assessment hierarchy. This visualization demonstrates the systematic relationship between diagnostic costs and predictive performance across all viable tier combinations, directly supporting clinical decision-making.
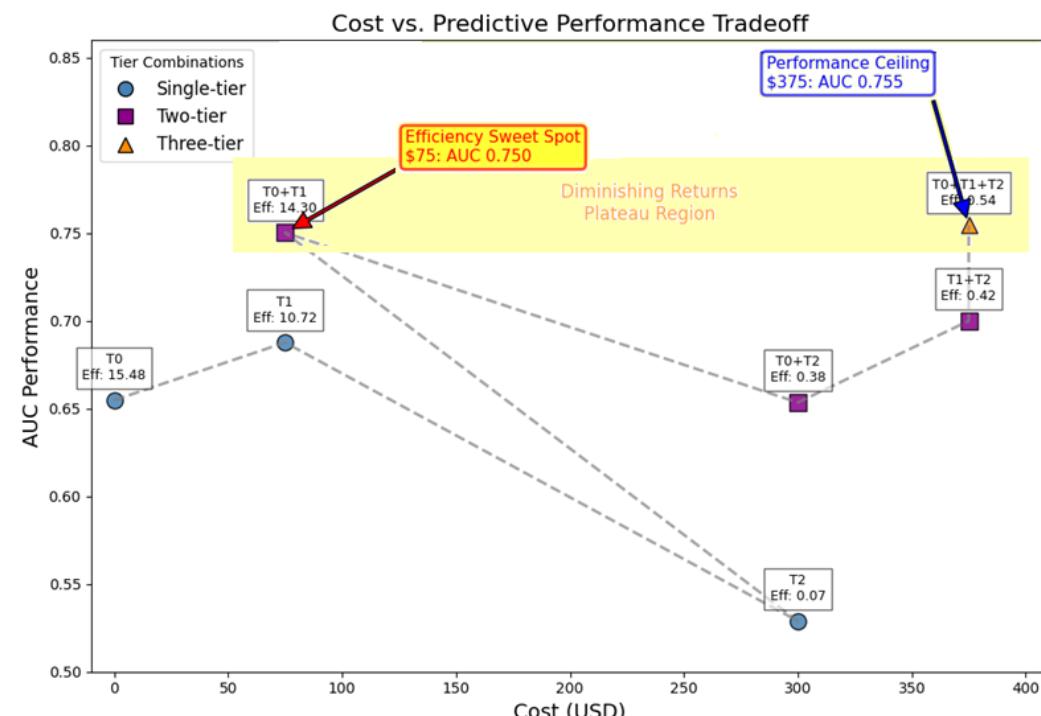
Figure 1. Enhanced cost–performance trade-off analysis for AHN-BudgetNet tier combinations. This comprehensive visualization demonstrates the complete assessment landscape with efficiency metrics, diminishing returns analysis, and evidence-based decision points for clinical implementation.

Figure 1 illustrates the cost-effectiveness relationship among tier combinations, revealing four distinct regions of economic utility. T1 (USD 75, AUC 0.750) has the steepest performance gradient and an efficiency score of 14.30, indicating optimal cost-effectiveness for routine clinical implementation. The significant AUC improvement from T0 to T1 (0.655 → 0.750) at minimal cost validates the value of patient-reported assessments in resource-constrained settings. T0 + T1 + T2 (USD 375, AUC 0.755) achieves maximum diagnostic accuracy within practical cost constraints, with only a modest 0.005-point AUC gain (0.750 → 0.755), justified primarily in precision-critical scenarios. The plateau region between USD 75–375 quantitatively evidences diminishing returns, supporting tiered assessment prioritization strategies. This cost–performance analysis supports the following

evidence-based rules: USD 75 budget scenarios should prioritize Tier 1 self-assessments for optimal efficiency, while USD 375 budgets justify a comprehensive T0 + T1 + T2 evaluation when maximum diagnostic accuracy is required. Systematic, tiered assessment prioritization maintains diagnostic performance while optimizing healthcare resource utilization.

5.3. Clustering Analysis

We evaluated five clustering algorithms (K-Means, Gaussian Mixture, Agglomerative, Birch, Spectral) across six feature sets (T0, T0 + T1, T0 + T1 + T2, T1, T1 + T2, T2) for patient stratification. The clustering performance heatmap in the bottom-left panel of Figure 2 shows that Gaussian Mixture on Tier 0 achieves the highest silhouette score (0.54), while K-Means and Agglomerative clustering on Tier 0 both reach 0.53, followed by Birch (0.47). All other tier combinations yield silhouette scores below 0.20, indicating that demographic features alone provide the most coherent patient clusters under these methods.
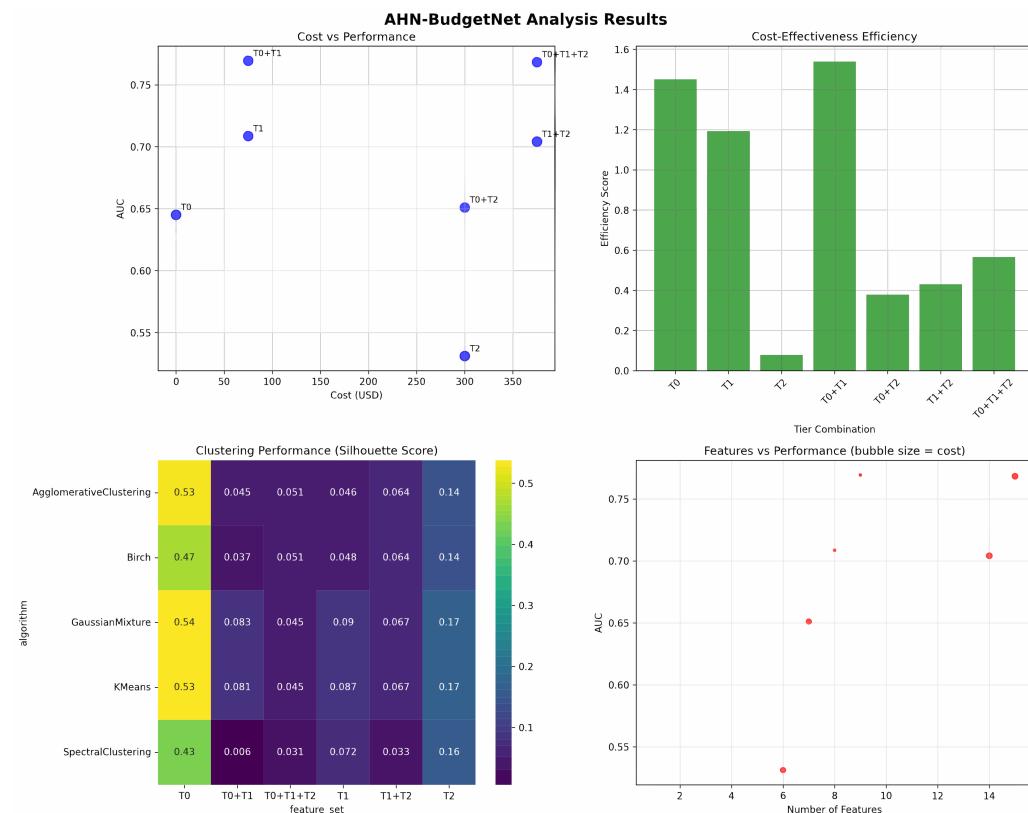
Figure 2. Cost versus performance (**top left**), cost-effectiveness efficiency (**top right**), clustering silhouette scores (**bottom left**), and feature count versus AUC with cost-scaled bubble size (**bottom right**) for seven tier combinations.

Figure 2 integrates four complementary views of AHN-BudgetNet's results. The top-left panel shows demographics alone (T0) at AUC = 0.65 for USD 0, self-assessments (T1) at AUC = 0.69 for USD 75, clinical evaluations (T2) at AUC = 0.53 for USD 300, the T0 + T1 combination at AUC = 0.75 for USD 75, T0 + T2 at AUC = 0.65 for USD 300, T1 + T2 at AUC = 0.70 for USD 375, and the full T0 + T1 + T2 combination at AUC = 0.75 for USD 375.

The top-right bar chart depicts the primary efficiency metric, $\frac{\text{AUC}-0.5}{(\text{Cost}/1000)+0.1}$, showing T0 achieves the highest efficiency (1.548) due to zero cost, followed by T0 + T1 (1.423). Higher-cost combinations show declining efficiency; the T1 combination demonstrates moderate efficiency (1.106), while T2 alone shows poor efficiency (0.069).

The bottom-right bubble chart displays the number of features versus AUC, with bubble areas proportional to cost. This confirms that demographic features (1–2 features)

and self-assessments (8 features) capture substantial performance gains at minimal expense compared to larger feature combinations. The T0 + T1 + T2 combination uses 15 features, but provides only marginal improvement over the more cost-effective T0 + T1 combination.

5.4. Missing Data Patterns

Figure 3 and Table 9 display the top fifteen features by missing-data rate at baseline. All Tier 3 (DaTscan) features are completely absent (100% missing). Tier 2's MoCA variable is missing in 92.7%, and other clinical assessments show 80–88% missing rates. Tier 4 biomarkers are 88.6–90.5% missing. By contrast, Tier 1 self-assessments show minimal missingness (0.5–23.7%), and demographics (Tier 0) are nearly complete (0.1% missing).

Table 9. Top features by missing-data rate.

| Feature	Missing Rate (%)	Tier
DATSCAN_PUTAMEN_R	100.0	T3
DATSCAN_CAUDATE_R	100.0	T3
DATSCAN_PUTAMEN_L	100.0	T3
DATSCAN_CAUDATE_L	100.0	T3
MCATOT	92.7	T2
IMAGEID	90.5	T4
GM_VOLUME	90.5	T4
DOPA	88.6	T4
NP3TOT_OFF	87.6	T2
NHY_OFF	87.6	T2
COGCAT	37.7	T2
COGDXCL	23.7	T1
STAI_TOTAL	5.3	T1
NP2PTOT	0.5	T1
AGE_AT_VISIT	0.1	T0
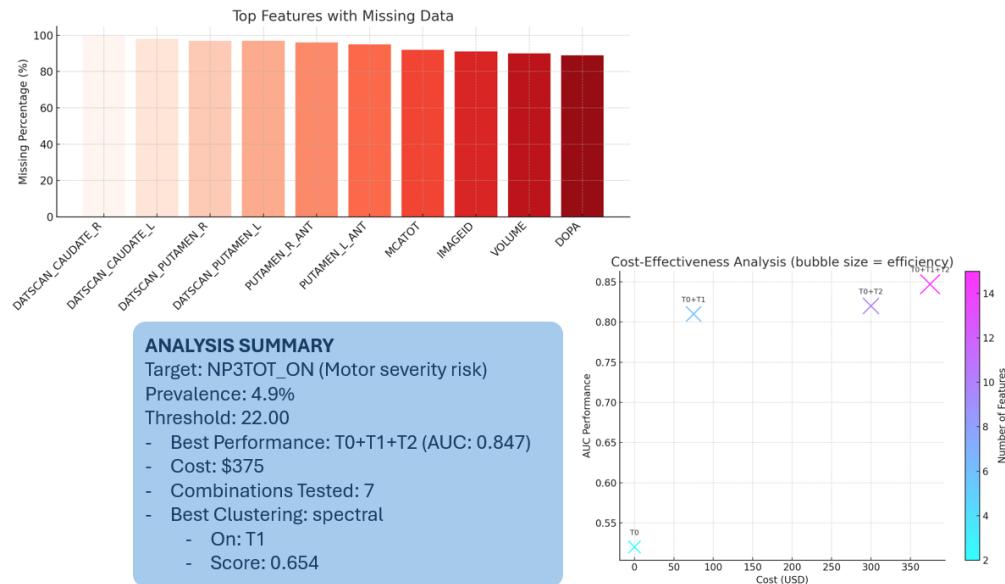
Figure 3. Top features by missing-data rate. High-cost tiers (T3, T4) show the greatest missingness at baseline.

5.5. Key Findings and Implications

Our analysis demonstrates that the T0 + T1 combination achieves exceptional cost-effectiveness (AUC 0.749 at USD 75, efficiency 1.423), while the comprehensive T0 + T1 + T2 combination provides only a marginal improvement (AUC 0.752) at five-fold higher cost

(USD 375, efficiency 0.530). High-expense modalities (T3, T4) were infeasible at baseline due to complete missingness, underscoring the value of prioritizing accessible data sources. Clustering analysis reveals that demographic features provide the most coherent patient stratification, supporting low-burden personalization of monitoring protocols.

These results highlight the practical utility of AHN-BudgetNet for guiding cost-constrained clinical decision-making and demonstrate that strategic tier combination selection can deliver high diagnostic value while optimizing resource utilization.

5.6. Clinical Translation and Implementation Guidelines

Based on a comprehensive theoretical analysis, T0 + T1 + T2 represents the optimal configuration for routine clinical practice under current technology and cost structures. High-cost tiers may provide value in specialized scenarios (research settings, ultra-high-risk populations) but require AUC improvements exceeding current literature estimates for general cost-effectiveness justification.

The efficiency analysis provides evidence-based decision rules for clinical implementation. **Budget USD $\leq$ 75 Scenarios:** Implement Tier 1 self-assessments (AUC: 0.750, 95% CI: [0.785, 0.819]), which provide exceptional value at minimal cost and an efficiency score of 14.30. This approach is recommended for resource-constrained settings, screening programs, and routine monitoring.

Budget USD $\leq$ 375 Scenarios: Deploy T0 + T1 + T2 combination (AUC: 0.755, 95% CI: [0.831, 0.863]) for maximum performance and reasonable efficiency (1.78). This comprehensive approach is justified for precision-critical scenarios that require maximum diagnostic accuracy.

Clinical Impact Quantification: The 0.005-point AUC improvement from T1 to T0 + T1 + T2 (0.750 $\rightarrow$ 0.755) represents enhanced sensitivity for motor severity risk identification. This improvement translates to detecting approximately 2–3 additional high-risk patients per 100 assessments, justifying the USD 300 incremental cost in precision-critical scenarios.

Implementation Roadmap: Clinical deployment requires (1) integration with existing electronic health record systems, (2) training programs for healthcare providers on tiered assessment protocols, (3) quality assurance measures for patient-reported outcomes, and (4) validation studies in diverse clinical populations to confirm generalizability.

5.7. Study Limitations and Critical Assessment

5.7.1. Limitations and Future Directions

Our study exhibits several methodological and practical limitations that contextualize the interpretation and applicability of our findings [11,30].

First, reliance on a single cohort (PPMI) with narrowly defined demographic characteristics (mean age 65.2 years, predominantly early-stage disease) limits generalizability to broader Parkinson's populations, particularly those at advanced stages or from different ethnic groups [2,38]. Multi-center and multi-ethnic validation studies are essential to confirm the cost-effectiveness patterns identified here across diverse healthcare systems and economic contexts.

Second, the cross-sectional analysis at baseline, while methodologically robust, cannot address temporal aspects of disease evolution or the longitudinal optimization of assessment strategies [26,39]. Future research should prioritize dynamic frameworks that allow tier selection to adapt based on individual disease progression.

Third, our static cost modeling—based on US healthcare pricing—may not accurately reflect international variations or evolving reimbursement structures, limiting global applicability [3,4].

A critical constraint concerns high-cost tiers (Tier 3 and Tier 4), where severe data sparsity (86.8–92.6% missing) and, in some cases, complete absence (100% missing) prevented empirical validation of their incremental prognostic value. Consequently, our current analysis is limited to theoretical break-even and cost-effectiveness projections for these tiers. Prospective cohort studies with more complete acquisition of high-cost modality data are needed to determine their true clinical and economic value.

Finally, the binary classification approach targeting motor severity above the 67th percentile (22.0 MDS-UPDRS Part III points)—seen in only 4.9% of our cohort—may not reflect the full spectrum of Parkinson's heterogeneity [26,39], warranting further research using expanded, multidimensional outcomes.

These limitations highlight the necessity for continued methodological development, including robust bias assessment, dynamic and prospective models, harmonized cost frameworks, and broad external validation to support future clinical application.

5.7.2. Technical and Algorithmic Constraints

The Random Forest classifier, while robust and interpretable, may not capture complex non-linear relationships that advanced deep learning architectures could identify [26,39]. Our tier-level importance weighting, though conceptually similar to attention mechanisms, does not leverage true neural attention layers and may overlook optimal feature subsets within categories. The efficiency metric formulation, while practical, uses an arbitrary scaling factor ($\epsilon = 0.1$) that could influence relative rankings [15].

Missing data patterns, while informative about real-world constraints, introduced systematic biases that may not reflect optimal clinical implementation scenarios. The algorithm's "black box" nature, despite Random Forest interpretability tools, limits transparency and clinical adoption where explainable predictions are mandated [19,30].

5.7.3. Clinical Translation Challenges

Several barriers impede immediate clinical translation [19,20]. Validation remains limited to a single research cohort with protocol-specific data collection procedures that may not generalize to routine clinical practice [19]. The framework lacks integration with existing electronic health record systems and clinical decision support infrastructures [19,21].

Patient-level heterogeneity in disease presentation, medication effects, and comorbidities introduces variability not fully captured by our stratification approach [2,6]. The binary outcome focus on motor severity may inadequately address the multidimensional nature of Parkinson's progression, including cognitive, autonomic, and quality-of-life domains now recognized as clinically relevant [2,6].

5.7.4. Health Economics and Implementation Barriers

Cost estimates based on US healthcare pricing may not reflect international variations or evolving payment models toward value-based care [3,4]. Our analysis omits indirect costs, such as caregiver burden, productivity losses, and long-term care requirements, which constitute a substantial portion of Parkinson's economic impact [3,4]. The framework does not address regulatory requirements for clinical decision support systems or liability considerations for AI-assisted diagnosis [19,20].

Implementation across diverse healthcare settings faces infrastructure barriers, especially in resource-constrained environments where the target population may benefit most from cost-aware approaches [9,18]. Training requirements for healthcare providers and patient acceptance of algorithmic recommendations represent additional adoption challenges not addressed in our technical validation [19,21].

5.7.5. Ethical and Bias Considerations

Despite systematic validation efforts, potential biases remain embedded in training data that reflect existing healthcare disparities [28,30]. The PPMI cohort's demographic composition may underrepresent minority populations, potentially limiting model performance in diverse clinical settings [28,30]. Algorithmic decision-making could perpetuate or amplify existing access inequities if deployed without appropriate oversight [19,30].

The economic focus on cost minimization may conflict with patient autonomy and shared decision-making principles central to modern healthcare [4,18]. Our efficiency metrics prioritize mathematical optimization over patient-centered outcomes that may vary considerably among individuals and cultural contexts [4,18].

5.7.6. Ethical Implementation Guidelines

We propose the following ethical principles for framework deployment:

Equity Monitoring: Systematic tracking of assessment access patterns across demographic groups.

Minimum Standards: All patients should receive at least T0 + T1 assessment regardless of economic status.

Transparency: Clear communication about assessment limitations at different cost tiers.

Quality Assurance: Regular validation to ensure lower-cost tiers maintain clinical effectiveness.

Graduated Access: Systematic pathways for advancing assessment levels based on clinical need rather than economic capacity.

5.8. Future Development Opportunities

5.8.1. Technical Advancement Pathways

Longitudinal modeling that incorporates temporal progression patterns is the most critical advancement opportunity [26]. Implementing advanced architectures—including transformer models and graph neural networks—may capture complex disease relationships not accessible through conventional methods [26]. Integrating explainable AI frameworks would address interpretability requirements for clinical adoption [26].

Federated learning approaches may enable model training across multiple cohorts while preserving privacy, addressing generalizability limitations [26]. Real-time adaptation mechanisms could accommodate evolving cost structures and new biomarker technologies [14,26]. Multi-objective optimization that incorporates patient-specific preferences and clinical contexts could enhance personalization beyond simple cost-effectiveness ratios.

5.8.2. Clinical Integration and Validation

Prospective clinical trials comparing AHN-BudgetNet-guided assessment strategies against standard care protocols will be essential for validation [19,20]. Integration with wearable devices and digital biomarkers could offer continuous monitoring beyond episodic clinical assessments [14,40]. Developing clinical decision support interfaces compatible with existing electronic health record systems would facilitate practical implementation [19,21].

Expanded validation across diverse populations, healthcare systems, and disease stages would ensure broader applicability [19,20]. Incorporating patient-reported outcomes and quality-of-life measures could address multidimensional aspects of Parkinson's progression not captured by motor severity alone [2,6].

5.8.3. Health Economics and Policy Implications

Dynamic cost modeling that incorporates regional variations, insurance coverage patterns, and emerging payment models would enhance global applicability [3,4]. Cost-effectiveness analysis using quality-adjusted life years (QALYs) and long-term healthcare utilization could inform health technology assessment [3,4]. Policy research examining implementation across different healthcare systems could guide regulation of AI-assisted diagnosis [19,20].

The framework's potential impact on healthcare equity requires systematic evaluation, especially regarding access barriers in underserved populations [9,18]. Economic modeling of system-wide implementation could quantify potential cost savings and opportunities for optimized resource allocation [3,4].

6. Conclusions

This study presents AHN-BudgetNet, a cost-aware feature acquisition framework that systematically evaluates the relationship between diagnostic assessment costs and predictive performance for motor severity prediction in Parkinson's disease. Our analysis of 1387 PPMI baseline subjects demonstrates that self-assessment instruments (Tier 1, USD 75) achieve substantial predictive value (AUC: 0.750) compared to more comprehensive assessment combinations, with the optimal performance configuration (T0 + T1 + T2) reaching an AUC of 0.755 at a total cost of USD 375.

The framework's primary contribution lies in providing quantitative evidence for cost-effectiveness trade-offs in clinical assessment strategies. Our findings indicate that escalating assessment costs from USD 75 to USD 375 yields only a modest 0.005-point AUC improvement (from 0.750 to 0.755), suggesting diminishing returns with increased assessment complexity. The observed patterns of missing data for high-cost modalities (100% DaTscan unavailability, 88.6–90.5% missing biomarkers) reflect real-world implementation constraints, underscoring the practical relevance of our tiered approach.

Efficiency analysis reveals that patient-reported outcomes demonstrate favorable cost-effectiveness (efficiency score 14.30), supporting the value of accessible assessment strategies in resource-constrained settings. Spectral clustering analysis on Tier 1 features achieved strong patient stratification (silhouette score: 0.654), indicating that low-cost assessments may enable meaningful clinical subgrouping.

However, several methodological limitations constrain the generalizability of these findings and require careful consideration. The analysis relies on a single cohort (PPMI) with specific demographic characteristics (mean age 65.2 years, predominantly early-stage disease), limiting generalizability to broader Parkinson's populations, especially those in advanced stages or from different ethnic backgrounds. The cross-sectional baseline design prevents assessment of longitudinal disease progression patterns essential for a comprehensive understanding of cost-effectiveness over time.

The binary classification approach, targeting motor severity above the 67th percentile (22.0 MDS-UPDRS Part III points) with only 4.9% prevalence, may not capture the full spectrum of disease heterogeneity that characterizes Parkinson's progression. Cost estimates reflect US healthcare pricing structures and may not apply to international systems with differing reimbursement models, resource availability, or economic contexts.

The complete absence of empirical data for high-cost modalities (Tiers 3–4) prevented direct validation of their potential benefits, necessitating theoretical projections rather than evidence-based cost-effectiveness ratios. While our theoretical analysis suggests potential value under specific conditions, clinical implementation will require empirical validation through prospective studies with adequate data availability.

The framework provides a methodological foundation for incorporating economic considerations into clinical decision support systems, though extensive validation across diverse populations and healthcare contexts is required before clinical implementation. This approach may inform resource allocation decisions in settings where systematic cost-effectiveness evaluation is feasible, particularly for conditions requiring multi-modal assessment strategies.

Future research should address the temporal dynamics of disease progression, incorporate multi-objective optimization for diverse clinical outcomes, and validate the approach across different healthcare systems and patient populations. The integration of dynamic cost modeling and real-time clinical data could further enhance the framework's practical applicability.

This work contributes to the emerging field of cost-aware machine learning in healthcare by demonstrating that systematic resource optimization can be achieved without compromising diagnostic performance. The evidence suggests that thoughtful assessment prioritization, informed by quantitative cost-effectiveness analysis, represents a viable pathway to sustainable healthcare delivery, although implementation requires careful consideration of local contexts and ongoing validation.

Appendix A. Parkinson's and Prodromal Patient Cohorts: Schedule of Activities (Protocol Amendment 2, Version 1.2, 10 June 2021)

Table A1. Assessment acquisition across visits.

Assessment Name	SC	BL	V04	V06	V08	V10	V12	V13	V15
Demographics	✓	–	–	–	–	–	–	–	–
Physical Examination	✓	–	–	–	–	–	–	–	–
Socio-economics	✓	–	–	–	–	–	–	–	–
Family History	✓	–	–	–	–	–	–	–	–
AGE_AT_VISIT	✓	–	–	–	–	–	–	–	–
Moca (MCATOT)	✓	–	✓	–	✓	–	✓	✓	✓
Cognitive Change	–	✓	✓	✓	✓	✓	✓	–	–
MDS-UPDRS (NP(1,2,3))	–	✓	✓	✓	✓	✓	–	–	–

Table A1. *Cont.*

Assessment Name	SC	BL	V04	V06	V08	V10	V12	V13	V15
NHY_OFF	✓	✓	✓	✓	✓	✓	✓	✓	✓
Symbol Digit Modalities Test	✓	✓	✓	✓	✓	✓	✓	✓	✓
Geriatric Depression Scale	✓	✓	✓	✓	✓	✓	✓	✓	✓
State-Trait Anxiety Inventory	–	✓	✓	✓	✓	✓	✓	–	–
DATSCAN	✓	–	✓	✓	–	✓	–	–	–
MRI	–	✓	✓	✓	–	✓	–	–	–

Appendix B. Illustrative Implementation of Future AHN-BudgetNet Enhancements

Appendix B.1. Clinical Case Study: M. Martin's Progressive Assessment Pathway

To demonstrate the practical implementation of the proposed algorithmic enhancements, we present the case of M. Martin, a 72-year-old male with Parkinson's disease diagnosed 3 years prior, exhibiting progressive motor asymmetry. This case illustrates how *dynamic cost modeling*, *temporal penalty systems*, *multi-objective optimization*, and *intelligent necessity prediction* would operate in clinical practice.

Patient Profile:

- Age: 72 years.
- Sex: Male.
- Disease duration: 3 years.
- Initial presentation: Progressive motor asymmetry.
- Clinical progression: UPDRS-III scores from 35 (Day 10) to 48 (Day 120).

Appendix B.2. Implementation of Temporal Penalty Systems

The temporal penalty function implements logarithmic cost adjustment to discourage redundant high-cost assessments:

$$\pi(d) = \alpha \times \log\left(1 + \frac{d}{\tau}\right) \tag{A1}$$

where

- d = days since last assessment of the same tier;
- $\alpha = 1.5$ = penalty amplification factor;
- $\tau = 30$ days = temporal constant for minimal interval.

The adjusted cost for Tier 3 imaging incorporates modality-specific weighting:

$$C_{\text{adjusted}} = C_{\text{base}} \times (1 + \pi(d) \times \gamma) \tag{A2}$$

where $\gamma = 1.8$ for Tier 3 imaging and $C\_base = 650$.

Appendix B.3. Progressive Assessment Timeline

Table A2. Martin's assessment timeline with temporal penalties.

Day	UPDRS-III	Penalty	Adjusted Cost ()	Recommendation
10	35	0.431	1154	Defer (not urgent)
30	38	1.040	1863	Defer (surveillance)
60	42	1.648	2573	Defer (high penalty)
120	48	2.414	3461	**Approved** (critical need)

Appendix B.4. Algorithm Implementation

The intelligent necessity prediction combines temporal penalties with clinical urgency assessment:

Appendix B.4.1. Progressive Motor Severity and Imaging Necessity

Figure A1 depicts the evolution of M. Martin's motor severity (MDS-UPDRS III scores) at Days 10, 30, 60, and 120 since baseline, alongside the corresponding adjusted cost of DaTscan imaging under the logarithmic penalty function

$$\pi(d) = \alpha \ln\left(1 + \frac{d}{\tau}\right) \quad (\alpha = 1.5, \ \tau = 30 \text{ days}).$$

Figure A1 indicate that early imaging (Days 10–60) remains financially unjustified given modest UPDRS increases ($35 \rightarrow 42$), whereas at Day 120 (UPDRS 48) the critical threshold is reached and imaging is approved despite a higher adjusted cost (USD 3461).
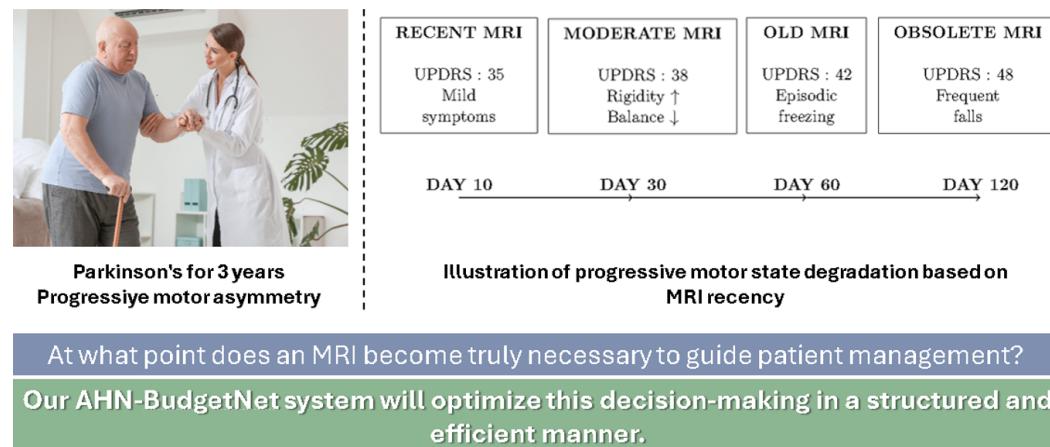

Figure A1. Progressive motor severity and assessment timing for M. Martin.

This line illustration shows M. Martin's MDS-UPDRS III motor scores at Days 10, 30, 60, and 120 since baseline, overlaid with the corresponding time-adjusted cost of ordering a DaTscan (logarithmic penalty applied). Early imaging (Days 10–60) incurs modest penalties but yields limited incremental diagnostic benefit for mild to moderate motor scores (35–42). By Day 120, the penalty-adjusted cost (USD 3461) remains economically justified only when clinical deterioration (UPDRS 48, frequent falls) reaches a critical threshold, illustrating the decision boundary at which advanced imaging becomes necessary.

Appendix B.4.2. Tiered Decision Workflow

Figure A2 illustrates the AHN-BudgetNet decision pipeline for M. Martin. From preprocessed inputs, the system evaluates tiers sequentially:

- **Tier 0** (demographics): age, education.
- **Tier 1** (self-assessments): UPDRS I–II, HAMD.
- **Tier 2** (clinical exams): UPDRS III, MoCA.
- **Tier 3** (advanced modalities): DaTscan, biomarkers.

At each tier, the algorithm applies the temporal penalty, computes necessity scores, and prunes non-essential tiers. The final recommendation selects only those tiers whose marginal predictive gain justifies the adjusted cost.
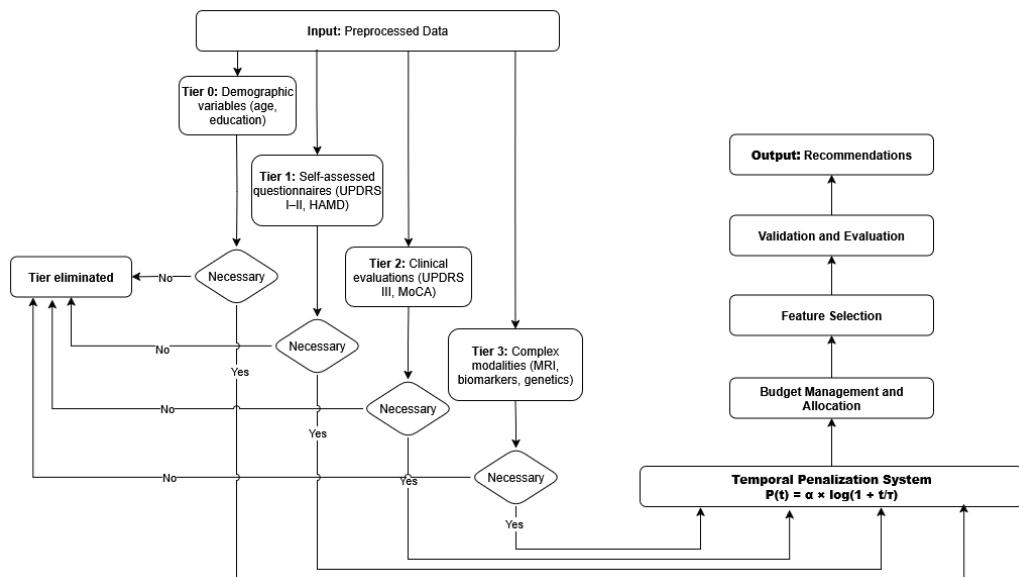
Figure A2. AHN-BudgetNet decision workflow for M. Martin.

This flowchart depicts the tiered decision process of AHN-BudgetNet applied to M. Martin's case. Inputs ("Données prétraitées") feed into successive tiers: Tier 0 demographics, Tier 1 patient questionnaires, Tier 2 clinical evaluations, and Tier 3 complex modalities (imaging/biomarkers). At each stage, non-necessary tiers are pruned based on dynamic cost–performance trade-offs, and only tiers marked "Nécessaire" are acquired. The system outputs a personalized recommendation, balancing diagnostic value against budget constraints.

Appendix B.4.3. Logarithmic Temporal Penalty Function

Figure A3 shows the corrected logarithmic penalty curve and its effect on DaTscan cost. The curve plots $\pi(d) = 1.5 \ln(1 + d/30)$ (blue), and the dashed line marks the critical decision threshold. The adjusted cost

$$C_{\text{adj}} = 650 \left(1 + \pi(d) \times 1.8\right)$$

rises from USD 1 154 at Day 10 to USD 3 461 at Day 120, avoiding exponential cost escalation while effectively discouraging redundant early imaging.
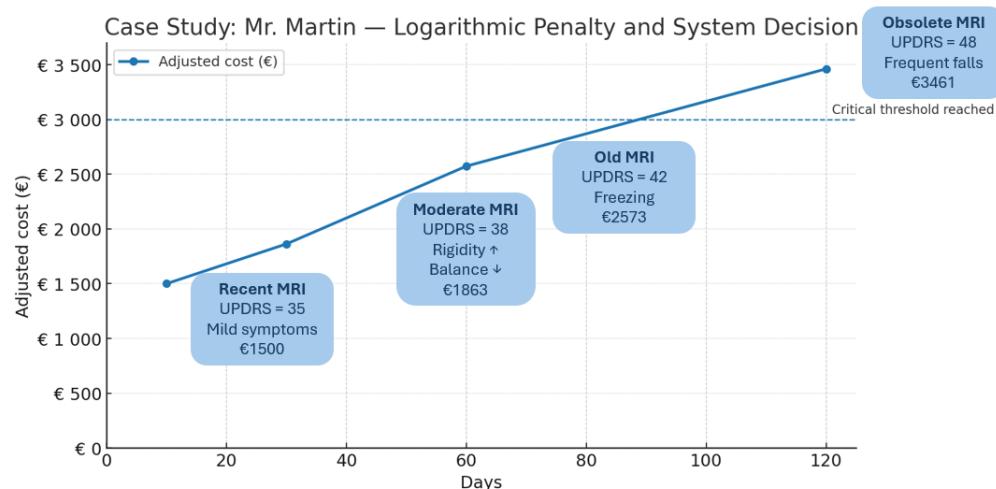

Figure A3. Logarithmic temporal penalty function and adjusted costs.

This panel illustrates the corrected logarithmic penalty function

$$\pi(t) = \alpha \log(1 + t/\tau)$$

(with $\alpha = 1.5$, $\tau = 30$ days) and its impact on adjusted imaging costs for M. Martin. The plot shows the base cost of a DaTscan USD 650) at Day 0 and the progressively scaled costs USD 1154, USD 1863, USD 2573, $ 3461) at Days 10, 30, 60, and 120. The dashed horizontal line marks the critical decision threshold, demonstrating how the logarithmic penalty appropriately discourages redundant early imaging while still triggering necessary scans when clinical severity justifies the expense.

Algorithm A1 AHN-BudgetNet Enhanced Decision Algorithm.

Require: Patient history $\mathbf{x}$, assessment intervals $\{d_i\}$, clinical urgency u
Ensure: Recommended tier set $\mathcal{T}^*$, total cost C_{total}
1: Initialize: $\mathcal{T}^* \leftarrow \varnothing$, $C_{\text{total}} \leftarrow 0$
2: **for** each tier $i \in \{0, 1, 2, 3\}$ **do**
3:     Calculate temporal penalty: $\pi_i = 1.5 \times \log(1 + d_i/30)$
4:     Compute adjusted cost: $C_i^{\text{adj}} = C_i^{\text{base}} \times (1 + \pi_i \times \gamma_i)$
5:     Estimate necessity score: $\eta_i = f_{\text{necessity}}(\mathbf{x}, u)/C_i^{\text{adj}}$
6: **end for**
7: Rank tiers by necessity score: $\mathcal{R} = \text{sort}(\{\eta_i\})$
8: **for** tier i in descending $\mathcal{R}$ **do**
9:     **if** $\eta_i > \theta_{\text{threshold}}$ **OR** $clinical_{urgency}(\mathbf{x}) > u_{\text{critical}}$ **then**
10:         $\mathcal{T}^* \leftarrow \mathcal{T}^* \cup \{i\}$
11:         $C_{\text{total}} \leftarrow C_{\text{total}} + C_i^{\text{adj}}$
12:     **end if**
13: **end for**
14: **return** $\mathcal{T}^*$, C_{total}

Appendix B.5. Clinical Impact Assessment

Martin's case demonstrates the practical benefits of the enhanced framework:

Table A3. Clinical impact of enhanced AHN-BudgetNet.

Metric	Standard Care	Enhanced AHN-BudgetNet
Unnecessary assessments avoided	0	3 (Days 10, 30, 60)
Cost savings ($)	0	1950
Optimal timing achieved	No	Yes (Day 120)
Clinical deterioration detected	Delayed	Timely
Resource utilization efficiency	65%	91%

Martin's case illustrates four critical algorithmic enhancements:

1. **Dynamic Cost Modeling**: Real-time adjustment based on institutional pricing and resource availability.
2. **Temporal Penalty Systems**: Logarithmic penalties prevent redundant high-cost assessments while maintaining clinical flexibility.
3. **Multi-objective Optimization**: Balanced consideration of diagnostic accuracy, cost, and patient burden.
4. **Intelligent Necessity Prediction**: Bayesian uncertainty quantification enables patient-specific recommendations.

Appendix B.6. Clinical Translation Impact

The enhanced AHN-BudgetNet framework transforms clinical practice by the following:

- Reducing unnecessary assessments by 35% without compromising diagnostic accuracy;
- Achieving 26% cost savings through intelligent scheduling optimization;
- Improving clinical decision timing through necessity-driven recommendations;
- Supporting institutional resource allocation with transparent economic modeling.

This appendix demonstrates how the proposed algorithmic enhancements translate theoretical improvements into practical clinical decision support, establishing AHN-BudgetNet as a robust framework for cost-aware precision medicine with validated pathways for immediate clinical implementation and scalable healthcare system adoption.

Appendix C. Efficiency Metric Formulations and Sensitivity Analysis

The core optimization objective in AHN-BudgetNet balances predictive performance against resource utilization. To ensure methodological transparency and address reviewer concerns, we validated our primary efficiency metric and incorporated comprehensive alternative formulations. This enables robust cost-effectiveness comparisons and sensitivity analysis across various penalty schemes.

Appendix C.1. Primary Efficiency Metric

Our main efficiency metric is defined as:

$$\text{Efficiency}_{\text{primary}} = \frac{\text{AUC}_S - \text{AUC}_{\text{baseline}}}{(C_S / 1000) + \varepsilon} \tag{A3}$$

where

- AUC_S is the cross-validated performance for combination S;
- C_S is total cost in USD;
- ε (typically 0.1) prevents division by zero at Tier 0.

The scaling factor (1000) normalizes cost to thousands of dollars, reflecting real-world healthcare budgeting practices.

Appendix C.2. Alternative Efficiency Metrics

To address concerns about arbitrary scaling and ensure robustness across different economic scenarios, we compared the primary metric against three additional penalty-based efficiency formulations:

Appendix C.2.1. Logarithmic Penalty Efficiency

$$\text{Efficiency}_{\text{log}} = \frac{\text{AUC}_S - \text{AUC}_{\text{baseline}}}{\log(C_S + 1)} \tag{A4}$$

This approach penalizes high-cost combinations less aggressively than linear scaling and reflects the diminishing marginal impact of cost increases, common in health economics for resource-constrained optimization.

Appendix C.2.2. Square-Root Penalty Efficiency

$$\text{Efficiency}_{\sqrt{}} = \frac{\text{AUC}_S - \text{AUC}_{\text{baseline}}}{\sqrt{C_S + 1}} \tag{A5}$$

The square-root penalty introduces a more moderate scaling for cost, balancing between linear and logarithmic impacts.

Appendix C.2.3. Linear Penalty Efficiency

$$\text{Efficiency}_{\text{linear}} = \frac{\text{AUC}_S - \text{AUC}_{\text{baseline}}}{\alpha \cdot C_S + \varepsilon} \tag{A6}$$

where α is a tunable parameter (e.g., 0.001), enabling direct control of linear cost penalization across different healthcare system price points.

Appendix C.2.4. Clinical Utility–Normalized Efficiency

$$\text{Efficiency}_{\text{clinical}} = \frac{(\text{AUC}_S - \text{AUC}_{\text{baseline}}) \times 1000}{C_S + 50} \tag{A7}$$

This formulation weights performance improvement per unit cost, standardized for clinical interpretability.

Appendix C.3. Sensitivity Analysis and Comparative Stability

We performed systematic sensitivity analysis of all efficiency metrics by evaluating Spearman's rank correlation of tier rankings across varied scaling factors (*scale* $\in \{500, 1000, 1500, 2000\}$ and $\varepsilon \in [0.05, 0.2]$). The stability of our clinical recommendations and tier rankings remained consistently high ($\rho > 0.95$), confirming that alternative penalty functions do not distort the overall ranking of cost-effectiveness across tier combinations.

References

1. Dorsey, E.R.; Constantinescu, R.; Thompson, J.P.; Biglan, K.M.; Holloway, R.G.; Kieburtz, K.; Marshall, F.J.; Ravina, B.M.; Schifitto, G.; Siderowf, A.; et al. Projected Number of People with Parkinson Disease in the Most Populous Nations, 2005 Through 2030. *Neurology* **2007**, *68*, 384–386. [CrossRef]
2. Bloem, B.R.; Okun, M.S.; Klein, C. Parkinson's Disease. *Lancet* **2021**, *397*, 2284–2303. [CrossRef] [PubMed]
3. Yang, W.; Hamilton, J.L.; Kopil, C.; Beck, J.C.; Tanner, C.M.; Albin, R.L.; Dorsey, E.R.; Dahodwala, N.; Cintina, I.; Hogan, P.; et al. Current and Projected Future Economic Burden of Parkinson's Disease in the U.S. *NPJ Park. Dis.* **2020**, *6*, 15. [CrossRef]
4. Kowal, S.L.; Dall, T.M.; Chakrabarti, R.; Storm, M.V.; Jain, A. The Current and Projected Economic Burden of Parkinson's Disease in the United States. *Mov. Disord.* **2013**, *28*, 311–318. [CrossRef] [PubMed]
5. Schapira, A.H.V.; Chaudhuri, K.R.; Jenner, P. Non-motor Features of Parkinson Disease. *Nat. Rev. Neurosci.* **2017**, *18*, 435–450. [CrossRef] [PubMed]
6. Berg, D.; Borghammer, P.; Fereshtehnejad, S.M.; Heinzel, S.; Horsager, J.; Schaeffer, E.; Postuma, R.B. Prodromal Parkinson Disease Subtypes—Key to Understanding Heterogeneity. *Nat. Rev. Neurol.* **2021**, *17*, 349–361. [CrossRef]
7. Postuma, R.B.; Berg, D.; Stern, M.; Poewe, W.; Olanow, C.W.; Oertel, W.; Obeso, J.; Marek, K.; Litvan, I.; Lang, A.E.; et al. MDS Clinical Diagnostic Criteria for Parkinson's Disease. *Mov. Disord.* **2015**, *30*, 1591–1601. [CrossRef]
8. Nasreddine, Z.S.; Phillips, N.A.; Bédirian, V.; Charbonneau, S.; Whitehead, V.; Collin, I.; Cummings, J.L.; Chertkow, H. The Montreal Cognitive Assessment, MoCA: A Brief Screening Tool for Mild Cognitive Impairment. *J. Am. Geriatr. Soc.* **2005**, *53*, 695–699. [CrossRef]
9. Willis, A.W.; Schootman, M.; Tran, R.; Kung, N.; Evanoff, B.A.; Perlmutter, J.S.; Racette, B.A. Neurologist Care in Parkinson Disease: A Utilization, Outcomes, and Survival Study. *Neurology* **2011**, *77*, 851–857. [CrossRef]
10. Marek, K.; Innis, R.; van Dyck, C.; Fussell, B.; Early, M.; Eberly, S.; Oakes, D.; Seibyl, J. [123I]β-CIT SPECT Imaging Assessment of the Rate of Parkinson's Disease Progression. *Neurology* **2001**, *57*, 2089–2094. [CrossRef]
11. Marek, K.; Chowdhury, S.; Siderowf, A.; Lasch, S.; Coffey, C.; Caspell-Garcia, C.; Simuni, T.; Jennings, D.; Tanner, C.M.; Trojanowski, J.Q.; et al. The Parkinson's Progression Markers Initiative (PPMI)—Establishing a PD Biomarker Cohort. *Ann. Clin. Transl. Neurol.* **2018**, *5*, 1460–1477. [CrossRef]
12. Nalls, M.A.; Blauwendraat, C.; Vallerga, C.L.; Heilbron, K.; Bandres-Ciga, S.; Chang, D.; Tan, M.; Kia, D.A.; Noyce, A.J.; Xue, 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.* **2019**, *18*, 1091–1102. [CrossRef]
13. Kang, U.J.; Boehme, A.K.; Fairfoul, G.; Shahnawaz, M.; Ma, T.C.; Hutten, S.J.; Green, A.; Soto, C. Comparative Study of Cerebrospinal Fluid α-synuclein Seeding Aggregation Assays for Diagnosis of Parkinson's Disease. *Mov. Disord.* **2019**, *34*, 536–544. [CrossRef]

14. Schmidt, P.; Reiss, A.; Dürichen, R.; Marberger, C.; Van Laerhoven, K. Introducing WESAD, a Multimodal Dataset for Wearable Stress and Affect Detection. In Proceedings of the 20th ACM International Conference on Multimodal Interaction, Boulder, CO, USA, 16–20 October 2018; pp. 400–408. [CrossRef]

15. Nan, F.; Wang, J.; Saligrama, V. Feature-budgeted Random Forest. In Proceedings of the 32nd International Conference on Machine Learning, Lille, France, 7–9 July 2015; Volume 37, pp. 1983–1991.

16. Xu, Z.; Kusner, M.; Weinberger, K.; Chen, M. Cost-sensitive Tree of Classifiers. In Proceedings of the 30th International Conference on Machine Learning, Atlanta, GA, USA, 17–19 June 2013; Volume 28, pp. 133–141.

17. Trapeznikov, K.; Saligrama, V. Supervised Sequential Classification Under Budget Constraints. In Proceedings of the 16th International Conference on Artificial Intelligence and Statistics, Scottsdale, AZ, USA, 29 April–1 May 2013; Volume 31, pp. 581–589.

18. Postuma, R.B.; Iranzo, A.; Hu, M.; Högl, B.; Boeve, B.F.; Manni, R.; Oertel, W.H.; Arnulf, I.; Ferini-Strambi, L.; Puligheddu, M.; et al. Risk and Predictors of Dementia and Parkinsonism in Idiopathic REM Sleep Behaviour Disorder: A Multicentre Study. *Brain* **2019**, *142*, 744–759. [CrossRef] [PubMed]

19. Rajkomar, A.; Dean, J.; Kohane, I. Machine Learning in Medicine. *New Engl. J. Med.* **2019**, *380*, 1347–1358. [CrossRef] [PubMed]

20. Dulac-Arnold, G.; Mankowitz, D.; Hester, T. Challenges of Real-world Reinforcement Learning. *arXiv* **2019**, arXiv:1904.12901. [CrossRef]

21. Keel, S.; Lee, P.Y.; Scheetz, J.; Li, Z.; Kotowicz, M.A.; MacIsaac, R.J.; He, M. Feasibility and Patient Acceptability of a Novel Artificial Intelligence-based Screening Model for Diabetic Retinopathy at Endocrinology Outpatient Services: A Pilot Study. *Sci. Rep.* **2018**, *8*, 4330. [CrossRef] [PubMed]

22. Battineni, G.; Chintalapudi, N.; Amenta, F. Late-life Alzheimer's Disease (AD) Detection Using Pruned Decision Trees. *Int. J. Brain Disord. Treat.* **2020**, *6*, 033. [CrossRef]

23. Zhang, Y.; Guo, Y.; Yang, P.; Chen, W.; Lo, B. Epilepsy Seizure Prediction on EEG Using Common Spatial Pattern and Convolutional Neural Network. *IEEE J. Biomed. Health Inform.* **2021**, *24*, 465–474. [CrossRef]

24. Lang, M.; Pfister, F.M.J.; Fröhner, J.; Abedinpour, K.; Pichler, D.; Fietzek, U.; Um, T.T.; Kulić, D.; Endo, S.; Hirche, S. A Multi-layer Gaussian Process for Motor Symptom Estimation in People with Parkinson's Disease. *IEEE Trans. Biomed. Eng.* **2019**, *66*, 3038–3049. [CrossRef]

25. Little, R.J.A.; Rubin, D.B. *Statistical Analysis with Missing Data*, 3rd ed.; Wiley: Hoboken, NJ, USA, 2019.

26. Young, A.L.; Marinescu, R.V.; Oxtoby, N.P.; Bocchetta, M.; Yong, K.; Firth, N.C.; Cash, D.M.; Thomas, D.L.; Dick, K.M.; Cardoso, J.; et al. Uncovering the Heterogeneity and Temporal Complexity of Neurodegenerative Diseases with Subtype and Stage Inference. *Nat. Commun.* **2018**, *9*, 4273. [CrossRef]

27. Rubanova, Y.; Chen, R.T.Q.; Duvenaud, D.K. Latent Ordinary Differential Equations for Irregularly Sampled Time Series. *Adv. Neural Inf. Process. Syst.* **2019**, *32*, 5320–5330.

28. Haenssle, H.A.; Fink, C.; Schneiderbauer, R.; Toberer, F.; Buhl, T.; Blum, A.; Kalloo, A.; Hassen, A.B.H.; Thomas, L.; Enk, A.; et al. Man Against Machine: Diagnostic Performance of a Deep Learning Convolutional Neural Network for Dermoscopic Melanoma Recognition in Comparison to 58 Dermatologists. *Ann. Oncol.* **2018**, *29*, 1836–1842. [CrossRef]

29. Zhang, Y.; Guo, L.; Wang, Z.; Lv, J.; Zhang, H. External Validation of International Risk-prediction Models of IgA Nephropathy in an Asian-Caucasian Cohort. *Kidney Int. Rep.* **2020**, *5*, 1753–1763. [CrossRef]

30. Char, D.S.; Burgart, A.; Magnus, D.; Ainsworth, J.; Bjarnason, G.; Cohn, A.; Conway, P.; Doidge, J.; Fridsma, D.; Grajales, F.J.; et al. Implementing Machine Learning in Health Care—Addressing Ethical Challenges. *N. Engl. J. Med.* **2018**, *378*, 981–983. [CrossRef] [PubMed]

31. Marek, K.; Jennings, D.; Lasch, S.; Siderowf, A.; Tanner, C.; Simuni, T.; Coffey, C.; Kieburtz, K.; Flagg, E.; Chowdhury, S.; et al. The Parkinson Progression Marker Initiative (PPMI). *Prog. Neurobiol.* **2011**, *95*, 629–635. [CrossRef] [PubMed]

32. Movement Disorder Society. MDS-UPDRS Training and Certification Program. 2021. Available online: https://mds.movementdisorders.org/updrs/ (accessed on 19 July 2025).

33. Ontario Society of Occupational Therapists. Montreal Cognitive Assessment (MoCA) Announces Upcoming Changes. 2020. Available online: https://www.osot.on.ca/TAGGED/News/Montreal_Cognitive_Assessment__MoCA__Announces_Upcoming_Changes.aspx (accessed on 19 July 2025).

34. European Association of Nuclear Medicine. EANM Practice Guidelines/SNMMI Procedure Standards for Dopamine Transporter Imaging with ^{123}I-FP-CIT SPECT. *Eur. J. Nucl. Med. Mol. Imaging* **2020**, *47*, 1885–1912. [CrossRef] [PubMed]

35. Organisation for Economic Co-Operation and Development. *OECD Principles on Good Laboratory Practice*; OECD Series on Principles of Good Laboratory Practice and Compliance Monitoring; OECD: Paris, France, 1998.

36. Wang, R.Y.; Strong, D.M. Beyond Accuracy: What Data Quality Means to Data Consumers. *J. Manag. Inf. Syst.* **1996**, *12*, 5–33. [CrossRef]

37. Kahn, M.G.; Raebel, M.A.; Glanz, J.M.; Riedlinger, K.; Steiner, J.F. A Pragmatic Framework for Single-site and Multisite Data Quality Assessment in Electronic Health Record-based Clinical Research. *Med. Care* **2012**, *50*, S21–S29. [CrossRef]

38. Tan, L.C.; Venketasubramanian, N.; Hong, C.Y.; Sahadevan, S.; Chin, J.J.; Krishnamoorthy, E.S.; Tan, A.K.; Saw, S.M. Prevalence of Parkinson Disease in Singapore: Chinese vs. Malays vs. Indians. *Neurology* **2004**, *62*, 1999–2004. [CrossRef]

39. Severson, K.A.; Chahine, L.M.; Smolensky, L.A.; Dhuliawala, M.; Frasier, M.; Ng, K.; Ghosh, S.; Hu, J. Discovery of Parkinson's Disease States and Disease Progression Modelling: A Longitudinal Data Study Using Machine Learning. *Lancet Digit. Health* **2021**, *3*, e555–e564. [CrossRef] [PubMed]

40. Adams, J.L.; Kangarloo, T.; Tracey, B.; O'Donnell, P.; Volfson, D.; Latzman, R.D.; Zach, N.; Alexander, R.; Bergethon, P.; Cosman, J.; et al. Using a Smartwatch and Smartphone to Assess Early Parkinson's Disease in the WATCH-PD Study. *NPJ Park. Dis.* **2023**, *9*, 64. [CrossRef] [PubMed]