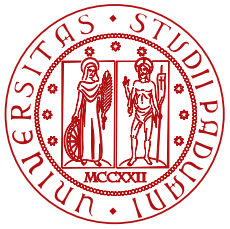


Università degli Studi di Padova

Department of Information Engineering
University of Padova

University of Padua
Department of Information Engineering
Master's Degree in Bioengineering

Impact of Batch Effects on Neuroimaging-Based Clinical Scales

Master Candidate

Mattia Longhini

Student ID 2123448

Supervisor

Prof. Mattia Veronese

University of Padua

Co-supervisor

Marco Pinamonti, Manuela Moretto

University of Padua

Academic Year: 2025/2026

Graduation Date: 23 April 2026

Abstract

Quantitative neuroimaging is redefining the role of magnetic resonance imaging in neurology by complementing visual assessment with objective and biologically meaningful markers of neurodegeneration. When these measures are translated into cutoff-based clinical scales, however, their value depends not only on biological sensitivity, but also on their ability to remain stable and interpretable across different acquisition and processing settings. Within this framework, the present thesis investigated whether batch effects and site-related variability may compromise the portability of neuroimaging-based clinical scales, using the imaging component of Stage 1 of the Huntington's Disease Integrated Staging System (HD-ISS) as a case study. The study was conducted on a large pooled multi-site structural MRI cohort processed through a common pipeline. Published HD-ISS thresholds were applied to an external normative sample; site effects were then assessed, different harmonization strategies were compared, and both external threshold recalibration and hierarchical normative modelling were explored. Direct application of the original thresholds revealed marked heterogeneity across datasets and a tendency to overestimate abnormality. Among the evaluated methods, ComBat proved the most effective in reducing site-related variability, although harmonization alone remained insufficient. The most robust solution emerged from the combination of harmonization and external recalibration, whereas normative modelling provided only partial correction. Overall, these findings highlight how crucial the management of technical variability is for making quantitative neuroimaging-based scales genuinely robust and transferable across multi-site settings.

Sommario

Il neuroimaging quantitativo sta ridefinendo il ruolo della risonanza magnetica in neurologia, affiancando alla valutazione visiva marker oggettivi e biologicamente significativi della neurodegenerazione. Quando però queste misure vengono tradotte in scale cliniche basate su soglie, il loro valore dipende non solo dalla sensibilità biologica, ma anche dalla capacità di rimanere stabili e interpretabili in contesti acquisitivi e analitici diversi. In questo quadro, la presente tesi ha indagato se batch effects e variabilità legata al sito possano compromettere la portabilità delle scale cliniche basate sul neuroimaging, utilizzando come caso di studio la componente di imaging dello Stage 1 dell'Huntington's Disease Integrated Staging System (HD-ISS). Lo studio è stato condotto su un'ampia coorte multi-site di MRI strutturale processata con una pipeline comune. Le soglie HD-ISS pubblicate sono state applicate a un campione normativo esterno; successivamente sono stati valutati gli effetti di sito, confrontati diversi approcci di armonizzazione ed esplorati sia la ricalibrazione esterna delle soglie sia il normative modelling gerarchico. L'applicazione diretta delle soglie originali ha mostrato una marcata eterogeneità tra dataset e una tendenza a sovrastimare l'anomalia. Tra i metodi valutati, ComBat si è dimostrato il più efficace nel ridurre la variabilità legata al sito, anche se la sola armonizzazione non è risultata sufficiente. La soluzione più solida è emersa dalla combinazione di armonizzazione e ricalibrazione esterna, mentre il normative modelling ha fornito solo una correzione parziale. Nel complesso, i risultati evidenziano quanto la gestione della variabilità tecnica sia decisiva per rendere le scale basate sul neuroimaging quantitativo realmente robuste e trasferibili in contesti multi-site.

Contents

1 Introduction

1.1 Context

Quantitative neuroimaging is progressively reshaping the role of magnetic resonance imaging in neurology, shifting it from a predominantly qualitative and visually interpreted technique toward the extraction of objective, reproducible, and biologically meaningful measurements [1]. Unlike conventional MRI, which mainly relies on relative signal intensities, quantitative approaches aim to derive parameters that can function as imaging biomarkers of tissue integrity, disease burden, and disease progression, thereby enabling a more precise and measurable characterization of pathology [1].

Among quantitative MRI applications, brain volumetry represents one of the most established and clinically accessible approaches [2, 3]. By quantifying regional and global changes in brain volume from structural MRI, volumetric analysis can reveal subtle patterns of atrophy that are often not reliably captured by visual inspection alone, and its value has been documented across several neurological disorders [2]. In particular, volumetric MRI has shown clinical relevance in conditions such as Alzheimer's disease, multiple sclerosis, and epilepsy, where quantitative structural measures can support diagnosis, improve disease characterization, and contribute to prognosis and therapeutic planning [2, 3].

However, the transition of quantitative neuroimaging from research to routine clinical use is not guaranteed by the availability of a measurable

biomarker alone [1, 3]. Its practical utility depends critically on methodological robustness, including standardized acquisition protocols, validated processing tools, reproducibility across scanners and sites, quality-control procedures, and clinically interpretable thresholds [1, 2]. This requirement becomes even more stringent when quantitative imaging measurements are incorporated into clinical scales or biological staging systems, because their clinical validity depends not only on sensitivity to pathology, but also on portability across sites and generalizability across heterogeneous technical settings [1, 4]. In this sense, scales based on quantitative neuroimaging represent a particularly relevant translational challenge. Their appeal lies in the possibility of converting continuous biological measurements into clinically interpretable thresholds, categories, or stages; however, this conversion also makes them especially vulnerable to technical variability. If the underlying biomarker is affected by site-specific acquisition differences, scanner effects, or processing-related variability, the resulting scale may lose transportability and become difficult to apply reliably across centers [1, 2]. Therefore, the clinical usefulness of quantitative neuroimaging-based scales ultimately depends on whether the corresponding measurements can remain stable and interpretable when transferred from the environment in which they were developed to independent real-world settings [3, 4].

1.2 Aims

Against this background, the present work aims to investigate whether batch effects and site-related variability may influence the usability of clinical scales based on quantitative neuroimaging measurements. More specifically, this thesis seeks to evaluate to what extent technical variability may compromise the transportability of imaging-derived thresholds across datasets, and to explore methodological strategies that may improve their robustness and portability.

1.3 Case study

As a case study, this thesis focuses on Huntington's disease, a genetically determined neurodegenerative disorder in which structural neuroimaging

biomarkers have shown particular promise for the identification of early neurodegeneration [5, 6]. Observational studies have shown that measurable brain changes, especially within the striatum, can be detected many years before traditional clinical diagnosis, and that caudate and putamen atrophy represent among the earliest and best validated imaging markers of disease progression [7, 8, 6].

This work specifically considers the Huntington's Disease Integrated Staging System (HD-ISS), a biologically grounded research framework that incorporates MRI-based caudate and putamen volumetry as landmark measures for early disease staging [4]. Because HD-ISS relies on quantitative volumetric thresholds to define early pathological stages, it provides a particularly relevant model for examining how technical variability, limited portability, and site-related effects may influence the practical usability of neuroimaging-based scales [4].

1.4 Methodology

This work is based on a large multi-site normative structural MRI cohort comprising more than 5,000 healthy participants from nine public datasets, all processed through a common pipeline. The methodological workflow was designed to assess whether quantitative neuroimaging-based cutoff scales remain applicable across heterogeneous acquisition sites, using the HD-ISS Stage 1 imaging framework as a case study.

The analysis was structured into four main steps:

- direct application of published HD-ISS age-dependent thresholds to TIV-normalized caudate and putamen volumes;
- statistical assessment of site effects and comparative evaluation of ComBat, ComBat-GAM, and CovBat harmonization strategies;
- definition of HD-ISS thresholds on harmonized normative data;
- exploration of hierarchical normative modelling as an alternative to explicit data harmonization.
- application of the best-performing pipeline to a smaller sample including healthy controls and HD subjects, to obtain an initial patient-level evaluation of its practical relevance.

Overall, the goal was to define a methodological pipeline to improve the portability of quantitative neuroimaging-based cutoff scales across datasets while preserving biologically meaningful variability.

1.5 Outline

This report is divided into the following six chapters:

1. The first chapter is this introduction to the work.
2. The second chapter presents the theoretical background of the thesis, including quantitative neuroimaging, structural MRI volumetry, Huntington's disease, the HD-ISS framework, and the main methodological concepts related to site effects, data harmonization, and normative modelling.
3. The third chapter describes the methodology adopted in this project. It introduces the multi-site normative MRI cohort, the preprocessing pipeline, the statistical assessment of site effects, the harmonization strategies considered, the retraining of HD-ISS thresholds, and the hierarchical normative modelling approach.
4. The fourth chapter reports the results obtained from the different stages of the analysis, including site-effect assessment, harmonization performance, recalibrated HD-ISS thresholds, and normative modelling findings.
5. The fifth chapter discusses the results in light of the aims of the thesis, with particular attention to the portability of quantitative neuroimaging-based cutoff scales across datasets and acquisition sites.
6. The final chapter summarizes the main conclusions of the work and outlines possible future developments.

2

Theoretical Background

This chapter presents the theoretical background necessary for a better understanding of the project's context. It encompasses the followings:

1. The first section provides a concise overview of quantitative neuroimaging and imaging biomarkers.
2. The second section introduces the biological and clinical framework of Huntington's disease.
3. The next section presents the Huntington's Disease Integrated Staging System (HD-ISS).
4. The fourth part discusses multi-site quantitative MRI and measurement variability.
5. The following section focuses on batch effects and statistical harmonization in multi-site neuroimaging.
6. The final section introduces normative models.

2.1 Quantitative neuroimaging and imaging biomarkers

Magnetic resonance imaging has long been central to the evaluation of neurological disorders because of its ability to provide high-resolution anatomical information in a safe and non-invasive manner. In routine practice, however, structural MRI is still interpreted predominantly on a qualitative basis, through the visual assessment of image contrast and gross anatomical abnormalities. Although this approach remains clinically indispensable, it is inherently more suited to the recognition of macroscopic changes than to the characterization of subtle or slowly evolving tissue alterations. Quantitative neuroimaging emerged in response to this limitation, with the aim of extracting objective measurements from MRI data that are more directly related to tissue structure and pathology [3, 1].

Within this framework, imaging biomarkers can be defined as image-derived quantitative characteristics that serve as indicators of normal biological processes, pathological processes, or response to therapeutic interventions. For these measures to be clinically meaningful, they should meet a number of key requirements:

- objective measurement on a defined scale;
- sensitivity to biologically relevant variation;
- adequate accuracy, precision, and reproducibility;
- sufficient standardization across acquisition and processing settings.

Their relevance therefore lies not only in biological sensitivity, but also in methodological robustness. Standardization of acquisition protocols, image processing workflows, and output interpretation becomes essential for the transition from descriptive neuroimaging to measurement-based inference [1].

2.1.1 Brain volumetry as a quantitative MRI application

Among the different applications of quantitative MRI, brain volumetry is one of the most established approaches for translating structural MRI

into measurable biomarkers. MRI-based volumetry quantifies global or regional brain volumes from structural images, typically through segmentation procedures that separate intracranial tissues into anatomically meaningful classes. In this sense, volumetric analysis does not replace visual inspection, but complements it by improving the detection of focal and subtle structural abnormalities that may be difficult to appreciate qualitatively, particularly when they develop gradually over time [3].

For this reason, brain volumetry has become increasingly relevant in both clinical neurology and neuroimaging research. By converting anatomical variation into objective quantitative measures, it can support disease characterization, longitudinal monitoring, and treatment evaluation across a broad range of neurological conditions. More broadly, its development reflects a fundamental shift in neuroimaging: structural brain changes are no longer interpreted only descriptively, but can also be measured, compared, and modelled as biologically meaningful markers of disease burden and progression [3, 4].

2.2 Huntington's disease

Huntington's disease (HD) is an autosomal dominant neurodegenerative disorder caused by an expanded CAG trinucleotide repeat in exon 1 of the *HTT* gene. The mutation leads to the production of mutant huntingtin, and the length of the CAG expansion is strongly associated with disease penetrance and, to a substantial extent, with the age at onset. In particular, CAG repeat lengths of 40 or more are generally considered fully penetrant, whereas alleles in the intermediate range of 36–39 repeats are associated with reduced penetrance [5, 4].

Clinically, HD is characterized by a progressive combination of motor, cognitive, and neuropsychiatric manifestations. Motor abnormalities include involuntary movements, impaired coordination, and other extrapyramidal signs; cognitive changes typically involve executive dysfunction and progressive decline; behavioural and psychiatric symptoms may include apathy, irritability, depression, and other disturbances of affect and motivation. Although these domains are often described separately, they are part of a single neurodegenerative process that unfolds over time with marked

inter-individual variability in symptom profile and rate of progression [5, 8].

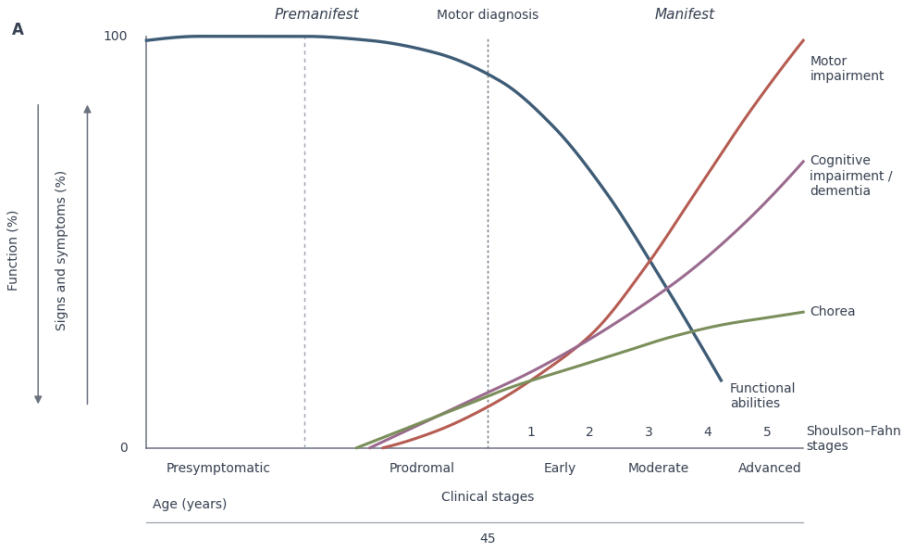
2.2.1 Premanifest and manifest disease: from biological onset to clinical diagnosis

A key feature of HD is that the biological disease process begins well before the stage at which a clinical diagnosis is traditionally made. In routine clinical and research settings, clinical motor diagnosis has commonly been operationalized through a diagnostic confidence level (DCL) of 4 on the Unified Huntington's Disease Rating Scale, indicating that the clinician is at least 99% confident that the observed motor abnormalities are attributable to HD. However, as emphasized by the HD-ISS framework, this diagnostic threshold reflects a relatively late point in the disease course rather than its true onset [4, 8].

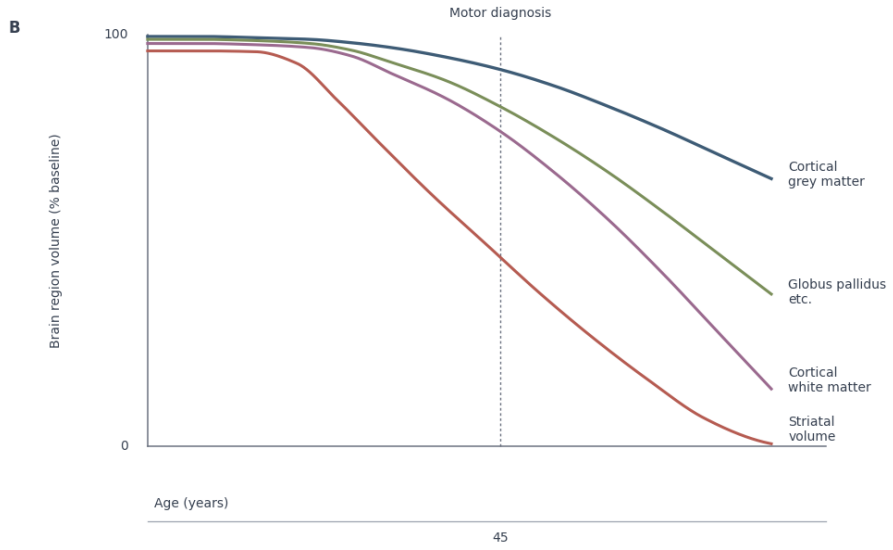
This interpretation is illustrated schematically in Figure 2.1a, where motor, cognitive, and functional changes evolve progressively across the premanifest and manifest phases rather than appearing abruptly at diagnosis. In the same conceptual framework, clinical motor diagnosis marks a transition within an ongoing pathological trajectory, not the beginning of disease itself. Accordingly, the traditional distinction between premanifest and manifest HD remains clinically useful, but it is biologically incomplete [5, 4].

A growing body of evidence supports this view by showing that structural, functional, and molecular alterations can be detected many years before clinical motor diagnosis. In particular, Figure 2.1b highlights the characteristic trajectory of regional brain volume loss across the disease continuum, showing that striatal atrophy emerges earlier and more prominently than more widespread cortical changes. This temporal dissociation between early biological change and later clinical diagnosis supports the view of HD as a continuous biological process rather than a dichotomous condition defined by phenoconversion [5, 4].

Evidence from very early premanifest cohorts further reinforces this perspective. In the HD-YAS study, young adult gene carriers approximately 24 years from predicted onset showed no detectable motor, cognitive, or psychiatric impairment despite the presence of measurable biological abnormalities, suggesting that pathological processes may already be underway while clinical function is still preserved [9].



(a) Schematic trajectory of clinical manifestations and functional decline across the HD continuum.



(b) Schematic trajectory of regional brain volume loss, highlighting the early and marked reduction in striatal volume.

Figure 2.1. Conceptual representation of Huntington's disease as a biological-clinical continuum. (A) Clinical and functional changes evolve progressively across premanifest and manifest phases. (B) Regional brain volume loss follows a characteristic pattern, with striatal atrophy preceding more widespread structural changes. Adapted from Ross et al. (2014).

Taken together, and as summarized in Figure 2.1, these findings support a framework in which HD is understood as a longitudinal continuum extending from genetic carriage to overt clinical and functional decline [9, 4,

5].

2.2.2 Structural MRI and striatal volumetric biomarkers in HD

Among the neuroimaging modalities investigated in Huntington's disease (HD), structural MRI is the most extensively studied and the most mature for biomarker research. Its widespread availability, non-invasive nature, absence of ionizing radiation, and high spatial resolution make it particularly suitable for longitudinal studies, especially in disease phases in which clinical changes remain subtle and conventional clinical measures may show limited sensitivity [5, 6].

Within this framework, the striatum has emerged as the most informative anatomical target. As shown in Figure 2.2, the striatum comprises the caudate nucleus and putamen, which are the structures most consistently affected by early HD-related neurodegeneration. Both neuropathological and imaging studies indicate that striatal atrophy precedes more widespread cortical involvement and can be detected many years before clinical motor diagnosis [5, 8, 6].

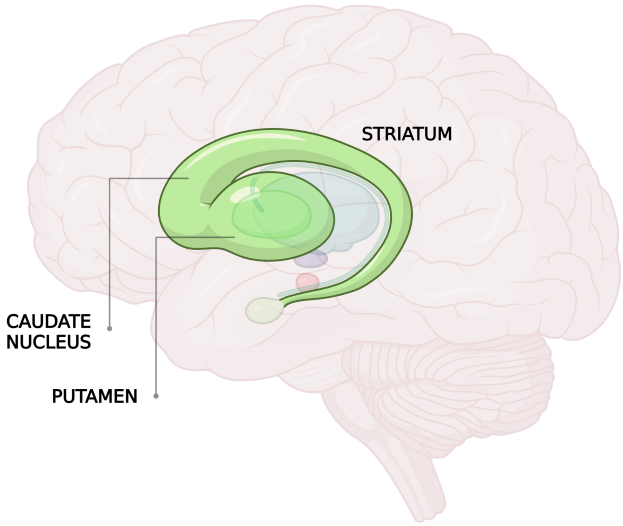


Figure 2.2. Schematic anatomical representation of the striatum within the human brain. The striatum, which includes the caudate nucleus and putamen, is a central target of early neurodegeneration in Huntington's disease.

Accordingly, caudate and putamen volumes have become the most robust structural MRI biomarkers of early disease progression. Cross-sectional

studies have shown reduced striatal volumes already in premanifest mutation carriers, with smaller volumes associated with increasing disease burden and closer proximity to clinical motor diagnosis. Longitudinal studies further demonstrated that these measures change reliably over time and provide some of the strongest imaging effect sizes in the prediagnostic phase, making them the best validated volumetric biomarkers currently available for use before clinical motor diagnosis [8, 6].

This evidence is also reflected in the Huntington's Disease Integrated Staging System (HD-ISS), in which head-size-adjusted caudate and putamen volumetric MRI were selected as the Stage 1 imaging landmarks because they were the biomarkers that showed detectable atrophy earliest in natural history modelling [4].

2.2.3 Evidence supporting caudate and putamen as volumetric biomarkers

Cross-sectional studies have repeatedly shown that caudate and putamen volumes are already reduced in premanifest gene carriers, with smaller striatal volumes associated with increasing disease burden and proximity to clinical motor diagnosis. In large observational cohorts, the relationship between striatal volume loss and disease progression appears broadly monotonic, with progressively greater atrophy as individuals approach or enter the clinically manifest phase. This makes the caudate and putamen particularly informative as quantitative indicators of underlying disease burden before overt clinical diagnosis [8].

Longitudinal evidence further supports their role. In their review, Kinnunen et al. identified MRI-based volumetric measures of the caudate and putamen as the best validated neuroimaging biomarkers for use before clinical motor diagnosis in Huntington's disease [8]. Across longitudinal studies, these measures showed robust change over time, with effect sizes that were generally larger in early manifest disease but already detectable in premanifest cohorts. The same body of evidence also indicates that striatal volume loss is sensitive to disease progression and sufficiently robust to remain a promising outcome measure for clinical trial applications [8, 6].

Prognostic value and relevance for early clinical research

The value of striatal volumetry lies not only in its sensitivity to anatomical change, but also in its prognostic relevance. Evidence reviewed indicates that baseline caudate and putamen volumes improve the prediction of clinical progression beyond age and CAG repeat length, and that putamen volume is among the strongest imaging predictors of time to clinical motor diagnosis [8]. Earlier work from PREDICT-HD and TRACK-HD similarly showed that baseline striatal measures can enhance prognostic accuracy and identify individuals at higher risk of near-term phenoconversion or motor deterioration [8].

Taken together, these findings explain why caudate and putamen volumetry has assumed a central role in early HD research. Structural MRI provides a technically accessible platform, and striatal volumetry offers a biologically grounded and empirically supported measure of early neurodegeneration. This combination makes the caudate and putamen especially suitable for frameworks that aim to identify HD before traditional clinical diagnosis, including biomarker-based staging systems such as the HD-ISS [4, 5, 6].

2.3 The Huntington's Disease Integrated Staging System (HD-ISS)

Traditional HD classification has largely relied on overt clinical manifestations, especially motor diagnosis. However, as discussed above, this threshold captures a relatively late point in the disease course. The HD-ISS was introduced to provide a research classification system that stages HD across its biological continuum by integrating genetics, biomarkers, clinical manifestations, and functional decline rather than symptoms alone [4].

Before introducing the staging structure itself, it is important to note that the HD-ISS also includes a biological case definition based on CAG repeat length and penetrance. Individuals with 40 or more CAG repeats are considered to meet the biological definition of HD, whereas for those with 36–39 repeats CAG length alone is not considered sufficient, because penetrance is reduced and additional disease-specific evidence is required.

For this reason, the operational staging framework proposed in the original HD-ISS paper primarily focuses on individuals with ≥ 40 CAG repeats [4].

2.3.1 Overview of stages 0–3

The HD-ISS divides HD progression into four cumulative stages. Stage 0 includes individuals who meet the genetic definition of HD but do not yet show detectable pathological, clinical, or functional change. Stage 1 marks the appearance of measurable biomarker evidence of pathogenesis. Stage 2 corresponds to the emergence of detectable clinical signs or symptoms, whereas Stage 3 indicates measurable functional decline. Because the system is cumulative, each stage includes the defining features of the previous ones, reflecting the progressive layering of biological, clinical, and functional change over time [4].

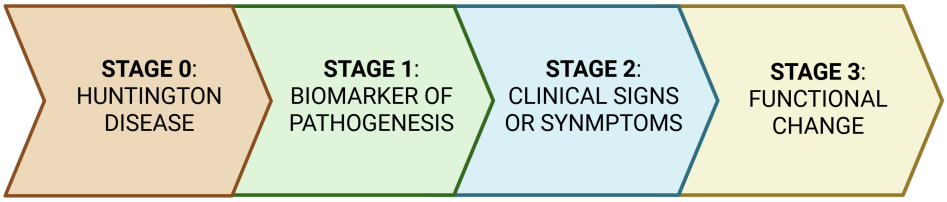


Figure 2.3. Overview of the Huntington's Disease Integrated Staging System (HD-ISS). The framework conceptualizes HD progression as a cumulative sequence from the genetic definition of disease to biomarker evidence of pathogenesis, detectable clinical manifestations, and functional decline.

Stage 1 as imaging-based evidence of neurodegeneration

Stage 1 is particularly relevant in the present thesis because it operationalizes the earliest measurable evidence of HD-related neurodegeneration. In the HD-ISS, this stage is defined by imaging landmarks derived from volumetric MRI of the caudate and putamen, adjusted for head size. These measures were selected because, in natural history modelling, they showed detectable abnormalities earliest and provided strong support as biomarkers of early disease progression [4].

Table 2.1. Summary of the Huntington's Disease Integrated Staging System (HD-ISS).

Stage	Domain	Landmark(s)	Interpretation
0	Genetic definition	Pathogenic CAG expansion	Biological definition of HD without detectable pathological, clinical, or functional change.
1	Biomarkers of pathogenesis	Caudate volume, putamen volume	Earliest measurable evidence of neurodegeneration.
2	Clinical signs or symptoms	Total Motor Score, Symbol Digit Modalities Test	Detectable clinical manifestation of disease.
3	Functional change	Total Functional Capacity, Independence Scale	Detectable decline in everyday function.

Landmark logic and age-dependent cut-offs

Stage assignment in the HD-ISS is based on landmark assessments evaluated against age-specific reference distributions derived from controls. For Stage 1, age-dependent cut-offs for caudate and putamen volumes were estimated from the control population, and the pathological threshold was defined as the extreme 5% tail of the normal distribution for age. These thresholds are therefore age-dependent but not CAG-dependent. Operationally, Stage 1 follows an either-or logic: crossing the cut-off for either the caudate or the putamen is sufficient to satisfy the stage criterion. This landmark-based approach is central to the HD-ISS because it translates continuous biological variation into a reproducible staging framework and directly underpins the methodological analyses developed in this thesis [4].

2.4 Multi-site quantitative MRI and measurement variability

A major challenge in the use of quantitative MRI biomarkers across large datasets is that measured values may reflect not only biological variation, but also technical differences related to image acquisition. In multi-site studies, scanner manufacturer, field strength, acquisition protocol, hardware configuration, and processing pipeline can all introduce systematic shifts in imaging measurements. These non-biological sources of variability, often referred to as scanner or site effects, may inflate variance, reduce statistical power, and generate findings that are difficult to compare or reproduce across datasets [10].

This issue is particularly relevant when quantitative imaging measures are used as biomarker candidates. Unlike qualitative image interpretation, which remains partly anchored to visual clinical judgement, quantitative measures are expected to be comparable across subjects, sites, and studies. Their usefulness therefore depends not only on biological sensitivity, but also on methodological consistency. If site-related variability is not adequately controlled, differences in measured brain volumes may partly reflect acquisition conditions rather than underlying neurobiology, thereby limiting the interpretability and portability of the biomarker itself [10, 1].

2.4.1 Harmonization and biomarker portability across sites

In this context, harmonization has emerged as a key strategy for reducing non-biological variability while preserving the effects of interest associated with biological covariates. Among the available approaches, ComBat has become one of the most widely adopted methods for multi-site neuroimaging data because it models and removes scanner-related effects while retaining variation associated with factors such as age, sex, or diagnosis. Originally developed in genomics and later adapted to neuroimaging, ComBat has been shown to improve consistency across scanners and to enhance the reproducibility of downstream analyses [10].

For biomarker-based frameworks such as the HD-ISS, this issue is especially important. Stage 1 classification relies on quantitative MRI measures of

the caudate and putamen, and therefore assumes that these measurements are sufficiently comparable across acquisition settings. In multi-site studies, however, technical variability may alter the apparent position of an individual relative to the expected reference distribution, potentially affecting the application of age-dependent cut-offs. Harmonization is therefore not only a statistical refinement, but also a methodological prerequisite for evaluating whether imaging-based staging criteria can be transferred reliably across datasets acquired under different technical conditions. This consideration directly motivates the methodological analyses developed in the following chapters.

2.5 Batch effects and statistical harmonization in multi-site neuroimaging

In multi-site neuroimaging, *batch effects* refer to unwanted sources of variation that do not reflect underlying biological differences but instead arise from differences in data acquisition or processing. In MRI studies, *scanner effects* more specifically indicate non-biological variability associated with scanner hardware or acquisition settings, whereas *site effects* can be understood more broadly as variability related to the imaging center, potentially including scanner-related as well as center-specific procedural factors [10, 11]. In this sense, site effects represent a particular form of batch effect in which observed measurements are partly shaped by the acquisition environment rather than exclusively by biology.

Non-biological variability in MRI can arise from several sources, including scanner manufacturer, magnetic field strength, gradient and radiofrequency coil characteristics, sequence parameters, image reconstruction procedures, subject positioning, and longitudinal scanner drift [10] . Additional heterogeneity may also be introduced downstream by preprocessing pipelines, software versions, and quality-control procedures. In pooled multi-site datasets, site effects may further incorporate differences in recruitment strategy, sample composition, and inclusion or exclusion criteria, so that "site" becomes a composite source of both technical and cohort-related variability [11, 12].

These effects are methodologically relevant because they can shift feature

Table 2.2. Main sources of site-related variability in multi-site neuroimaging studies, adapted from Bayer et al. (2022)[11].

Domain	Category	Examples
Technical site-related factors	Scanner platform	Contrast-to-noise ratio
		Coil inhomogeneity
	Sequence	Sensitivity to tissue differences
		Sensitivity to motion
	Acquisition procedure	Task details
		Instructions
		Circadian effects
		Motion/immobilisation
	Image post-processing	Residual variability despite pipeline harmonization
Neurobiological and sample-related factors	Demographic factors	Age
		Sex/gender
		Background population
	Clinical characteristics	Severity of the condition of interest
		Age-by-disease interaction
		Neurological and medical comorbidity
		Past psychiatric disease
	Treatment-related factors	Past medication
		Current medication
	Sampling and study design	Definition of healthy controls
		Clinical inclusion criteria
		Diagnostic instruments
		Study protocol differences

distributions, alter mean and variance, and remain detectable even when acquisition protocols are nominally harmonized [10]. As a consequence, they may obscure genuine biological associations, generate spurious findings, and reduce the reproducibility and comparability of downstream analyses. Their impact can be even more problematic in multivariate settings, where residual site information may remain strong enough to support accurate site or scanner classification from imaging features, indicating that technical variability can meaningfully structure the data beyond the biological signal of interest [13, 12].

2.5.1 Statistical representation and detection of site effects

At a general level, an imaging feature can be represented as the combination of a biological component of interest, a systematic site-related com-

ponent, and a residual error term. A simple formulation is:

$$Y_i = \mu + B_i + S_i + \varepsilon_i$$

where Y_i is the observed value for subject i, μ is the overall mean, B_i denotes the biological contribution, S_i the site-related contribution, and ε_i the residual error. This decomposition provides the conceptual basis for later harmonization models, whose main goal is to reduce S_i while preserving B_i [10, 11].

Before harmonization, site effects can be assessed using an ANCOVA model in which site is treated as a categorical factor and biological covariates such as age and sex are included as adjustment terms:

$$Y_i = \beta_0 + \beta_1 \text{Age}_i + \beta_2 \text{Sex}_i + \sum_{s=1}^{S-1} \gamma_s I(\text{Site}_i = s) + \varepsilon_i$$

where the coefficients γ_s test whether site explains variability beyond the biological covariates. In multi-site MRI studies, ANCOVA is therefore useful as a preliminary diagnostic framework to establish whether a measurable site effect is present before applying harmonization methods [11].

Statistical significance alone, however, is not sufficient to describe the magnitude of site effects. For this reason, variance-explained measures such as partial η^2 are useful to quantify how much of the total variability is attributable to site after accounting for biological covariates. This makes effect-size estimates informative both before harmonization, to characterize the severity of the problem, and after harmonization, to quantify the amount of residual site variance that remains [12].

2.5.2 Statistical harmonization techniques

Statistical harmonization aims to reduce unwanted site-related variability while preserving the biological content of the data, so that measurements become more comparable across acquisition settings without losing subject-level information of interest [11]. In operational terms, a harmonization method should attenuate non-biological differences across sites while retaining meaningful associations with biological covariates such as age or sex [10].

The main harmonization targets help distinguish the available approaches. Classical ComBat is designed to remove site-related shifts in *mean* and *variance* through location-and-scale adjustment [10]. CovBat extends this idea by addressing residual differences in *covariance*, which may remain relevant in multivariate and machine-learning settings [13]. ComBat-GAM retains the same harmonization logic but replaces linear covariate modeling with a more flexible formulation to preserve *nonlinear biological effects*, especially age-related trajectories across the lifespan [14, 11].

For this reason, preservation of biological signal is a central theoretical constraint of harmonization rather than a secondary property. A method that removes site effects but also attenuates genuine biological variation is not methodologically desirable, particularly when biological covariates are unevenly distributed across sites and nuisance variance may partly overlap with meaningful signal.

ComBat

ComBat is a statistical harmonization method originally developed for gene expression data and later adapted to neuroimaging to remove site-related variability while preserving biological effects of interest [15]. Its key idea is to model site effects through a *location-and-scale adjustment*, assuming that each feature may be affected by both an additive shift and a multiplicative scaling factor that vary across sites. This extends simple regression-based residualization by explicitly accounting for site-related differences in both mean and variance, while using Empirical Bayes to stabilize parameter estimation, especially in small samples [10, 11].

Using the notation adopted in the neuroimaging formulation of ComBat, the model for feature v measured in subject j from site i can be written as:

$$y_{ijv} = \alpha_v + X_{ij}^T \beta_v + \gamma_{iv} + \delta_{iv} \varepsilon_{ijv}$$

where y_{ijv} is the observed value of feature v for subject j at site i, α_v is the overall intercept for feature v, $X_{ij}^T \beta_v$ represents the contribution of the biological covariates to be preserved, γ_{iv} is the additive site effect, δ_{iv} is the multiplicative site effect, and ε_{ijv} is the residual error term, assumed to have mean zero [10].

After estimating the site parameters, the ComBat-harmonized values are

obtained by removing the additive and multiplicative site components and then reintroducing the preserved biological signal:

$$y_{ijv}^{\text{ComBat}} = \frac{y_{ijv} - \hat{\alpha}_v - X_{ij}\hat{\beta}_v - \gamma_{iv}^*}{\delta_{iv}^*} + \hat{\alpha}_v + X_{ij}\hat{\beta}_v$$

where γ_{iv}^* and δ_{iv}^* are the Empirical Bayes estimates of the site effects [10]. In practical terms, ComBat aims to suppress nuisance site variation while maintaining biologically meaningful associations, for example with age or sex. A key assumption is that covariate effects and site effects can be disentangled with sufficient support in the data; when biological covariates are strongly unbalanced across sites, correction becomes more delicate because nuisance variance and meaningful signal may partly overlap [10, 11].

CovBat

CovBat was introduced to extend ComBat beyond harmonization of mean and variance, addressing also site-related differences in the *covariance structure* of neuroimaging features, which can remain relevant in multivariate and machine-learning analyses [13, 11] . The method starts from ComBat-adjusted residuals and assumes that, even after removal of additive and multiplicative site effects, residual covariance matrices may still differ across sites [13].

After the initial ComBat step, the adjusted residuals for feature f, subject i, and site s can be written as:

$$e_{isf} = \frac{Y_{isf} - \hat{\alpha}_f - X_{is}^T\hat{\beta}_f - \hat{\gamma}_{sf}}{\hat{\delta}_{sf}}$$

where Y_{isf} is the observed value, $\hat{\alpha}_f$ is the intercept, $X_{is}^T\hat{\beta}_f$ represents the preserved biological covariates, and $\hat{\gamma}_{sf}$ and $\hat{\delta}_{sf}$ are the ComBat estimates of the additive and multiplicative site effects [11]. CovBat then performs principal component analysis on these residuals and harmonizes the principal component scores across sites, so that residual covariance differences are reduced in a lower-dimensional space [13, 11]. The site-adjusted score

for principal component k is expressed as:

$$\xi_{isk}^{\text{CovBat}} = \frac{\xi_{isk} - \hat{\mu}_{sk}}{\hat{\rho}_{sk}}$$

where ξ_{isk} is the original score for subject i, site s, and component k, while $\hat{\mu}_{sk}$ and $\hat{\rho}_{sk}$ are the site-specific location and scale parameters estimated in principal component space (Bayer et al., 2022). The adjusted residuals are then reconstructed in the original feature space and the preserved biological effects are added back:

$$y_{isf}^{\text{CovBat}} = e_{isf}^{\text{CovBat}} + \hat{\alpha}_f + X_{is}^T \hat{\beta}_f$$

In practical terms, CovBat retains the logic of ComBat but adds an explicit correction for covariance, making it particularly relevant when site effects influence the joint distribution of features rather than only their marginal mean or variance. A main limitation is that its performance may become less stable in smaller samples with many features, and unlike ComBat-GAM it does not incorporate nonlinear covariate modeling [13, 11].

ComBat-GAM

ComBat-GAM was introduced to extend the ComBat framework when the biological effects to be preserved cannot be adequately described by linear terms, especially in lifespan datasets where age-related trajectories are often nonlinear [14]. The method retains the location-and-scale adjustment of ComBat, but replaces the linear covariate component with a generalized additive model, allowing flexible smooth functions of the biological covariates while still correcting site-related differences in mean and variance [14].

In this formulation, the biological contribution is written as a smooth function:

$$f(X_{isf}) = \alpha + f_1(x_1) + f_2(x_2) + \cdots + f_n(x_n)$$

where $f(X_{isf})$ denotes the covariate-dependent biological component for feature f, α is the intercept, and $f_1, \ldots, f_n$ are smooth or linear terms associated with the covariates [11]. In the lifespan setting proposed, age is modeled nonlinearly, while sex and intracranial volume can be included as

linear terms [14]:

$$f(x_{\text{age}}, x_{\text{sex}}, x_{\text{ICV}}) = \alpha + f_1(x_{\text{age}}) + b_2 x_{\text{sex}} + b_3 x_{\text{ICV}}$$

After estimation of the smooth biological term, harmonization follows the same general logic as ComBat:

$$Y^*_{isf} = \frac{Y_{isf} - f(X_{isf}) - \hat{\gamma}_{sf}}{\hat{\delta}_{sf}} + f(X_{isf})$$

where Y_{isf} is the observed value for subject i, site s, and feature f, $f(X_{isf})$ is the preserved biological component, and $\hat{\gamma}_{sf}$ and $\hat{\delta}_{sf}$ are the estimated additive and multiplicative site effects [11]. In practical terms, ComBat-GAM is designed to suppress nuisance site variation while preserving nonlinear biological trends that would be oversimplified under a standard linear ComBat model. Its main advantage therefore lies in settings with broad age ranges, whereas in smaller or narrower datasets the added flexibility may be less stable and not necessarily beneficial [14].

2.5.3 Theoretical frameworks for harmonization assessment

A harmonization method should satisfy two fundamental criteria. First, residual site effects should be minimized, so that site explains as little variance as possible after correction and, ideally, is no longer easily identifiable from the harmonized data. Second, biological variability should be preserved, meaning that plausible associations with variables such as age or sex should remain detectable after correction. A method that suppresses site effects at the cost of attenuating genuine biological signal is not methodologically desirable [10, 11].

These criteria can be assessed using complementary strategies. One family of approaches is based on *effect-size estimates*, such as the proportion of variance attributable to site after adjustment for biological covariates. In this framework, statistics such as partial η^2 are useful because they quantify the residual magnitude of the site effect before and after harmonization. Another family is based on *prediction*, testing whether imaging site can still be inferred from the harmonized features. In recent work, harmonization efficacy was evaluated through site prediction accuracy compared both with

chance level and with the non-harmonized condition, showing that a reduction in site predictability provides a practical measure of residual unwanted variation [12].

In machine-learning settings, this evaluation must also consider *data leakage*. If harmonization parameters are estimated on the full dataset before train/test splitting, information from outside the training set may leak into the model-building process and lead to overly optimistic performance estimates. For this reason, harmonization should be fitted on training data only and then applied to unseen test data [12]. In the present work, an additional derived summary index, termed *Harmonization Effectiveness Score* (HES), was defined to summarize the proportional reduction of the site effect from ANCOVA-based effect sizes:

$$\mathrm{HES} = \max\left(0,\ 1 - \frac{\eta^2_{\mathrm{post}}}{\eta^2_{\mathrm{pre}}}\right)$$

This score is not a standard metric from the literature, but a project-specific derived index intended to provide a compact summary of harmonization performance.

Summary of the methodological landscape

Overall, batch effects are a structural feature of multi-site neuroimaging and require dedicated statistical treatment. Within this landscape, ComBat addresses site-related differences in mean and variance, CovBat extends correction to covariance structure, and ComBat-GAM adapts the same harmonization logic to nonlinear biological effects, particularly age-related trajectories across the lifespan.

2.6 Normative models

Normative models are statistical frameworks designed to estimate the expected distribution of a biological measure as a function of clinically relevant covariates, so that each individual can be evaluated relative to a reference population rather than against a group average. The classical analogy is that of pediatric growth charts: instead of comparing subjects only through mean differences between groups, normative modelling estimates centiles of variation and quantifies subject-specific deviations from the expected pattern [16, 17]. In neuroimaging, this means modelling structural or functional imaging features as a function of covariates such as age, sex, or site, and then expressing the discrepancy between observed and expected values at the level of the individual subject [16, 18].

In a probabilistic formulation, normative modelling estimates the predictive mean and variability of a response variable y conditional on a covariate matrix X:

$$y \sim \mathcal{N}(\mu, \sigma^2), \qquad \mu = f_\mu(X, \theta_\mu), \qquad \sigma = f_\sigma^+(X, \theta_\sigma)$$

and subject-level deviations can then be summarized as Z-scores [18]. In practical applications, a more complete deviation score also incorporates predictive noise and modelling uncertainty:

$$Z = \frac{y - \hat{y}}{\sqrt{\sigma^2 + \sigma_*^2}}$$

where $\hat{y}$ is the predictive mean, σ^2 is the predictive noise variance, and σ_*^2 is the uncertainty associated with model estimation [17, 19]. This probabilistic structure is especially useful in neuroimaging, where the aim is not only to predict the mean trend, but also to quantify how unusual an individual observation is relative to the normative range [16].

A key practical aspect is the separation between training and test data. The training set, ideally a healthy reference cohort, is used to estimate the normative model; the test set is then used to evaluate generalization out of sample and to compute deviation scores for unseen individuals [17]. In multi-site neuroimaging, this is particularly important because the objective is not to remove site effects in a preprocessing step, but to estimate a model

that can generalize while properly accounting for nuisance site variation inside the modeling framework itself [18, 11].

2.6.1 Hierarchical Bayesian regression

Hierarchical Bayesian regression (HBR) is one of the main probabilistic implementations of normative modelling for multi-site neuroimaging. Its central idea is that site-specific parameters are not estimated independently, but are assumed to arise from shared prior distributions. This implements partial pooling, which allows the model to borrow statistical strength across sites while still retaining site-specific structure [18, 11]. In other words, site effects are modelled rather than removed, making HBR a principled alternative to explicit harmonization.

In the multi-site formulation[18], the response in site i is written as

$$y_i = f_{\mu_i}(X, \theta_{\mu_i}) + \varepsilon_i$$

and the site-specific parameters are coupled through shared priors:

$$\theta_{\mu_i} \sim \mathcal{N}(\mu_{\theta_\mu}, \sigma^2_{\theta_\mu}), \qquad \theta_{\sigma_i} \sim \mathcal{N}(\mu_{\theta_\sigma}, \sigma^2_{\theta_\sigma})$$

Here, the lower-level parameters describe the relationship between covariates and imaging measures within each site, whereas the higher-level hyperparameters define how those site-specific parameters are distributed across sites. This hierarchy is the key mechanism by which HBR accommodates site variation without a separate harmonization step. From a Bayesian perspective, priors encode assumptions before observing the data, while posterior distributions update these assumptions after seeing the training data and are then used for prediction on test data or for recalibration on new sites [18, 11].

Gaussian HBR

In the Gaussian version of HBR, the response variable is assumed to follow a Gaussian likelihood. This is the simplest and most widely used formulation, and it is appropriate when the conditional distribution of the imaging phenotype is approximately symmetric and adequately summa-

rized by mean and variance. In this setting, the main parameters are the predictive mean and variance, while the priors are placed on the corresponding site-specific regression parameters and variance terms [18]. The Gaussian HBR is therefore attractive because it is fully probabilistic, naturally yields centiles and Z-scores, and separates aleatoric variability from epistemic uncertainty, which is central for subject-level inference[17].

SHASH HBR

A limitation of the Gaussian formulation is that many imaging-derived phenotypes are not well described by a normal distribution, because they may show skewness, kurtosis, or heteroscedasticity. To address this, de Boer et al. (2024) extended the HBR framework by replacing the Gaussian likelihood with a distribution from the sinh-arcsinh family, known as SHASH, while preserving the same hierarchical Bayesian logic for handling multi-site data. The general HBR formulation remains the same, but the likelihood is no longer Gaussian; instead, it is parameterized to explicitly control location, scale, skewness, and tail weight [20].

The SHASH distribution is obtained by applying an inverse sinh-arcsinh transformation to a standard Gaussian variable:

$$Z \sim \mathcal{N}(0,1), \qquad X = \xi_{\epsilon,\delta}^{-1}(Z), \qquad X \sim S(\epsilon,\delta)$$

with

$$\xi_{\epsilon,\delta}(x) = \sinh\big(\delta \operatorname{arcsinh}(x) - \epsilon\big)$$

In this family, ϵ governs skewness and δ governs tail weight and kurtosis. To make the parameters more interpretable and sampling more stable, it is introduced a reparameterized form [20], denoted $S_b(\mu,\sigma,\epsilon,\delta)$, in which μ and σ directly control location and scale, while ϵ and δ control asymmetry and shape. Under this reparameterization:

$$E[B] = \mu, \qquad \mathrm{Var}(B) = \sigma^2$$

This is important because it decouples location and scale from skewness and kurtosis, which improves parameter identifiability and MCMC sampling efficiency. In practical terms, SHASH HBR is preferable when the response distribution is visibly non-Gaussian or when the shape of the distribution

changes across the covariate range, for example across the lifespan [20].

In summary, both Gaussian and SHASH HBR follow the same conceptual strategy: they estimate a normative model on a reference training set, use hierarchical priors to model site-related variability rather than harmonizing it away, and then evaluate unseen test subjects through individualized deviation scores. The difference is that Gaussian HBR assumes a symmetric conditional distribution, whereas SHASH HBR extends the same framework to non-Gaussian responses with explicit modeling of skewness and kurtosis [20, 18].

3 Materials and Methodology

This chapter presents the materials and methodology employed in the present study. It first describes the multi-site structural MRI dataset, the the variables included in the analyses, and in the second part outlines the methodological framework adopted throughout the work.

3.1 Materials

This section describes both the pooled multi-site structural MRI resource used as the main normative dataset and the external iMarkHD-derived structural subset used in the study, together with their main acquisition characteristics, the final analytic samples derived from them, and the structural variables retained for subsequent analyses.

3.1.1 Pooled multi-site structural MRI resource

The present study was based on a pooled multi-site structural MRI resource assembled within the Department of Information Engineering (DEI), University of Padova. This resource was created by combining data from nine independent open neuroimaging cohorts and processing them through a common preprocessing workflow in order to obtain a comparable set of structural measures across datasets. Although the broader project also included resting-state functional MRI, only the structural component was con-

sidered in the present study.

Raw T1-weighted structural MRI scans first underwent image quality assessment with MRIQC [21] and were then processed within the fMRIPrep framework [22]. For the structural branch of the pipeline, preprocessing included intensity non-uniformity correction with N4BiasFieldCorrection, skull stripping using an ANTs-based brain extraction workflow, tissue segmentation into cerebrospinal fluid, white matter, and gray matter, and brain surface reconstruction with FreeSurfer 7.3.2 using *recon-all*. The workflow further included refinement of the brain mask and spatial normalization to standard MNI spaces, ultimately enabling the extraction of harmonized volumetric and cortical thickness measures across cohorts.

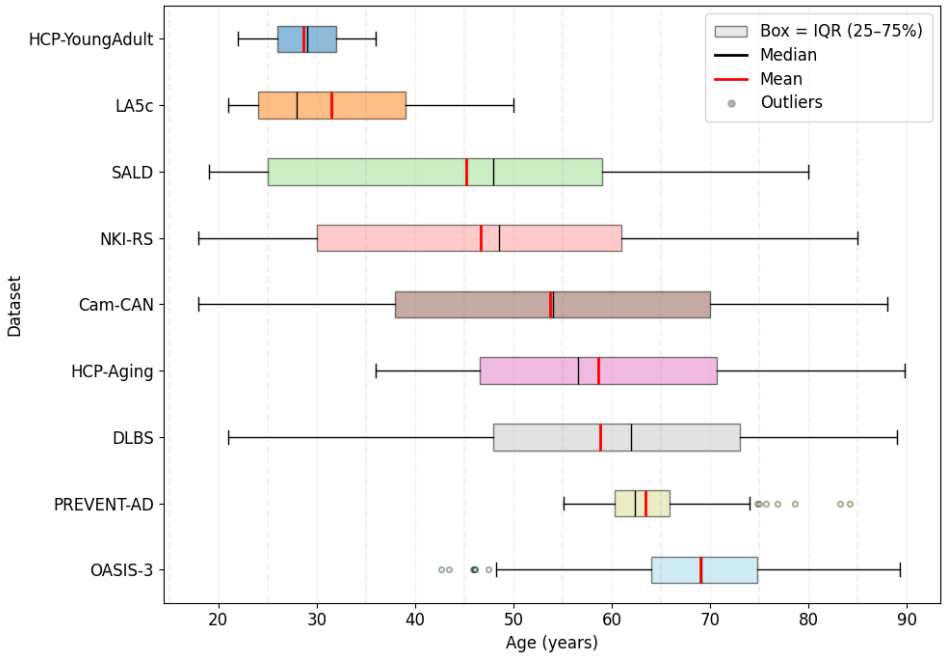
For the present study, the structural variables retained for analysis were dataset membership, age, sex, estimated total intracranial volume, total gray matter volume, lateral ventricle volume, hippocampal volume, caudate volume, and putamen volume. In addition, bilateral cortical thickness measures from the entorhinal, fusiform, inferior temporal, and middle temporal regions were retained to derive the temporal meta-ROI [23]. Among these variables, caudate and putamen volumes constituted the core structural measures of the present thesis, whereas the remaining variables provided complementary anatomical information and supported subsequent normalization and harmonization analyses.

Final analytic sample

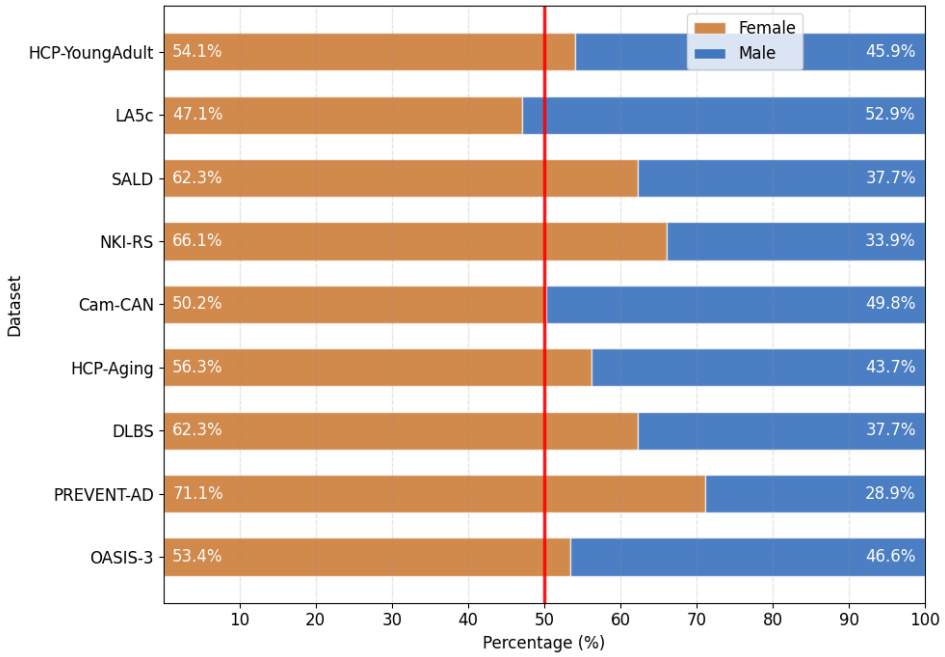
The final analytic sample was derived from the pooled structural MRI resource described above. Prior to analysis, participants older than 90 years were excluded, as their number was very limited and their inclusion could have introduced instability in age-dependent analyses. The resulting analytic sample comprised 5,084 participants across nine datasets.

The demographic composition of the final filtered sample is summarized in Table 3.1. For consistency throughout the manuscript, datasets were ordered according to the mean age of the final analytic sample in all subsequent descriptive summaries. The same ordering is reflected visually in Figure 3.1, where panel A (Figure 3.1a) shows the age distribution across datasets and panel B (Figure 3.1b) reports the corresponding sex distribution. All descriptive statistics, tables, and figures reported in this section therefore refer

to this final filtered sample.



(a) Age distribution by dataset.



(b) Sex distribution by dataset.

Figure 3.1. Demographic composition of the final analytic sample across datasets.

Table 3.1. Demographic characteristics of the final analytic sample across datasets. Participants older than 90 years were excluded prior to analysis. Dataset-specific references can be added in the first column or in a dedicated reference column, if preferred.

Dataset	N	Age range	Mean age ± SD	Female, n (%)	Male, n (%)
HCP-YoungAdult	1006	22.0–36.0	28.7 ± 3.7	544 (54.1%)	462 (45.9%)
LA5c	121	21.0–50.0	31.4 ± 8.7	57 (47.1%)	64 (52.9%)
SALD	491	19.0–80.0	45.2 ± 17.4	306 (62.3%)	185 (37.7%)
NKI-RS	894	18.0–85.0	46.7 ± 17.7	591 (66.1%)	303 (33.9%)
Cam-CAN	627	18.0–88.0	53.8 ± 18.5	315 (50.2%)	312 (49.8%)
HCP-Aging	599	36.0–89.8	58.6 ± 14.7	337 (56.3%)	262 (43.7%)
DLBS	414	21.0–89.0	58.8 ± 18.1	258 (62.3%)	156 (37.7%)
PREVENT-AD	301	55.1–84.2	63.4 ± 5.0	214 (71.1%)	87 (28.9%)
OASIS-3	631	42.7–89.3	69.0 ± 8.9	337 (53.4%)	294 (46.6%)
Total	5084	18.0–89.8	49.6 ± 19.2	2959 (58.2%)	2125 (41.8%)

3.1.2 Dataset and MRI acquisition characteristics

The contributing datasets differed in cohort design and imaging context, spanning healthy young-adult, adult lifespan, aging, and Alzheimer-related open neuroimaging resources. In line with the aims of the present study, only the structural MRI component of these datasets was considered.

An overview of the contributing cohorts and of the corresponding structural MRI acquisition context is reported in Table 3.2. As shown there, the pooled sample includes both single-site and multi-site resources, with variability in scanner manufacturer, platform, and protocol implementation. This heterogeneity was not incidental, but rather represented an important characteristic of the material, because it reflects the variability typically encountered when quantitative structural MRI measures are aggregated across independent neuroimaging resources.

Table 3.2. Characteristics of the contributing datasets and structural MRI acquisition protocols. Information was compiled from the official dataset documentation and related cohort publications.

HCP-YoungAdult [24, 25]	Healthy young adult cohort from the Human Connectome Project. Structural MRI data were acquired in a single-site setting on a customized Siemens Connectome Skyra 3T scanner using a high-resolution 3D T1-weighted MPRAGE sequence.
LA5c [26]	Open neuroimaging dataset including healthy controls and psychiatric groups. Structural scans were acquired across two sites on Siemens Trio 3T scanners using T1-weighted MPRAGE protocols.
SALD [27]	Southwest University Adult Lifespan Dataset. Structural MRI was acquired at a single site on a Siemens TrioTim 3T scanner using a sagittal 3D T1-weighted MPRAGE sequence.
NKI-RS [28]	Nathan Kline Institute Rockland Sample. Structural scans were acquired at a single site on a Siemens TIM Trio 3T scanner using a T1-weighted MPRAGE sequence.
Cam-CAN [29, 30]	Population-based adult lifespan cohort. Structural MRI data were acquired at a single site on a Siemens TIM Trio 3T scanner using a high-resolution 3D T1-weighted MPRAGE sequence.
HCP-Aging [31, 32]	Lifespan aging cohort from the Human Connectome Project. Structural MRI was acquired across multiple sites on Siemens Prisma 3T scanners using harmonized HCP-style T1-weighted protocols.
DLBS [33]	Dallas Lifespan Brain Study. Structural scans were acquired at a single site on a Philips Achieva 3T scanner using a 3D T1-weighted MP-RAGE sequence.
PREVENT-AD [34]	Cohort of older adults at risk for Alzheimer's disease due to family history. Structural MRI was acquired primarily on Siemens TIM Trio 3T systems using 3D T1-weighted MPRAGE sequences; protocol updates were introduced over time.
OASIS-3 [35]	Longitudinal aging and Alzheimer's disease resource. Structural MRI data were collected retrospectively across multiple Siemens scanners, with greater heterogeneity in platform and acquisition parameters; T1-weighted scans were predominantly MPRAGE-based.

Across cohorts, the analyses relied on measures derived from T1-weighted anatomical MRI. However, beyond the general sequence type, structural acquisition parameters also differed across datasets. For completeness, dataset-specific sequence parameters,including repetition time (TR), echo time (TE), inversion time (TI), flip angle, matrix dimensions, slice thickness, and voxel size, should be summarized in a dedicated table (Table 3.3) once all

protocol details have been collected from the original dataset documentation and source publications.

To facilitate traceability and reproducibility, the original reference for each dataset should also be reported explicitly, either directly in Table 3.2 or in an additional dedicated column. This will allow the reader to link each cohort description to its corresponding public release paper, dataset descriptor, or official documentation source.

Table 3.3. Dataset-specific structural MRI acquisition parameters for the T1-weighted sequences included in the final analytic sample. For OASIS-3, the reported values refer to a representative/common MPRAGE configuration rather than a unique protocol across the full dataset.

Dataset	TR (ms)	TE (ms)	TI (ms)	Flip angle (°)	Matrix dimensions	Voxel size / slice thickness
HCP-YoungAdult	2400	2.14	1000	8	$320 \times 320 \times 256$	$0.7 \times 0.7 \times 0.7$ mm
LA5c	1900	2.26	1100	7	$256 \times 256 \times 176$	$1 \times 1 \times 1$ mm
SALD	1900	2.52	900	9	$256 \times 256 \times 176$	$1 \times 1 \times 1$ mm
NKI-RS	1900	2.52	2000	9	$256 \times 256 \times 176$	$0.98 \times 0.98 \times 0.98$ mm
Cam-CAN	2250	2.99	900	9	$256 \times 240 \times 192$	$1 \times 1 \times 1$ mm
HCP-Aging	2500	1.8/3.6/5.4/7.2	1000	8	$320 \times 300 \times 208$	$0.8 \times 0.8 \times 0.8$ mm
DLBS	8.1	3.7	1100	12	$204 \times 256 \times 160$	$1 \times 1 \times 1$ mm
PREVENT-AD	2300	2.98	900	9	$256 \times 256 \times 176$	$1 \times 1 \times 1$ mm
OASIS-3	2400	3.08	1000	8	$256 \times 256 \times 176$	$1 \times 1 \times 1$ mm

3.1.3 Structural variables used in the present study

The present analyses focused on a subset of structural variables selected for their relevance to the study aims. These variables can be grouped as follows:

- **Demographic and dataset variables:** subject's identificative code, age, sex and dataset membership label.
- **Whole-brain volumetric variables:** estimated total intracranial volume and total gray matter volume.
- **Subcortical volumetric variables:** lateral ventricle volume, hippocampal volume, caudate volume, and putamen volume.
- **Temporal cortical thickness variables:** bilateral entorhinal, fusiform, inferior temporal, and middle temporal thickness.

Among the selected variables, caudate and putamen volumes constituted the central structural measures of the present work, as they served as the primary neuroanatomical markers throughout the main analyses.

Nevertheless, the remaining variables were also methodologically relevant. Estimated total intracranial volume was required to normalize the volumetric regions of interest, whereas biological covariates such as age and sex were essential for performing harmonization while preserving meaningful inter-individual variability. The dataset label was equally important, as it identified the acquisition site or cohort membership underlying site-related effects. In addition, the other volumetric and cortical thickness measures provided a broader structural framework that proved useful for comparing and evaluating different harmonization strategies in the subsequent stages of the study.

3.1.4 iMarkHD structural MRI subset

The present analysis was based on a structural MRI subset derived from the *iMarkHD* study, an observational longitudinal imaging study conducted at King's College London to investigate neuroimaging biomarkers of Huntington's disease progression. In the original study design, cohort 2 included healthy controls and participants with Huntington's disease across different clinical stages, classified into groups B, C, D, and E. For the purposes of the present work, only the structural MRI component was considered, and the analytic file was used as a derived cross-sectional volumetric dataset. According to the original study classification [36]:

- Group B comprised participants in an early neurodegenerative or preclinical symptomatic transition phase;
- Groups C and D represented more advanced neurodegenerative stages;
- Group E included matched healthy controls;

for the present analyses, groups C and D were combined into a single neurodegeneration category. MRI data were acquired at King's College London using a General Electric Discovery MR750 3T scanner equipped with a 32-channel head coil. The structural protocol included a high-resolution T1-weighted MPRAGE sequence with prospective motion correction, which served as the anatomical scan for subsequent structural analyses [36]. Volumetric measurements in the dataset made available for the present study were subsequently confirmed to have been derived using FreeSurfer 7.3.2 version.

For the present thesis, the retained variables were:

- subject identifier;
- age and sex;
- clinical group;
- estimated total intracranial volume;
- caudate volume;
- putamen volume.

In addition, normalized caudate and putamen measures were computed by dividing each regional volume by the estimated total intracranial volume, in order to account for inter-individual differences in head size. These normalized measures constituted the main structural variables used for comparison with HD-ISS volumetric thresholds.

After exclusion of subjects with missing demographic information, the final analytic sample comprised 92 participants. Group composition is summarized in Table 3.4.

Table 3.4. Composition of the iMarkHD structural MRI subset included in the present study.

Group	Description	N
E	Healthy controls	34
B	Early neurodegeneration / preclinical symptomatic transition	21
C+D	Neurodegeneration	37

3.2 Methodology

This section describes the analytical workflow used to assess the applicability of quantitative structural neuroimaging biomarkers within the HD-ISS framework in a large multi-site normative MRI cohort. First, the published HD-ISS Stage 1 thresholds were directly applied to TIV-normalized caudate and putamen volumes to evaluate their portability across datasets. Subsequently, site-related variability was formally assessed and different harmonisation strategies were compared. Finally, two complementary approaches were investigated to improve robustness across sites: re-estimation of HD-ISS thresholds on harmonised data and hierarchical normative modelling without explicit data harmonisation.

3.2.1 Application of published HD-ISS Stage 1 thresholds

As an initial step, the published HD-ISS Stage 1 imaging thresholds were directly applied to a large multi-site normative structural MRI cohort to assess whether they remained applicable outside the original development setting. In the HD-ISS framework, Stage 1 is defined by biomarker evidence of pathogenesis, and caudate and putamen volumetric MRI adjusted for head size were selected as the relevant imaging landmarks because they are the structural biomarkers that show atrophy earliest in disease progression. The corresponding thresholds are age-dependent cutoffs derived from the control distribution and defined at the 5th percentile, thereby identifying subjects whose volumes fall below the lower 5% of the age-matched normative distribution. Stage 1 eligibility is then evaluated according to the original either-or rule, whereby the imaging condition is satisfied when either putamen or caudate volume is below the corresponding age-specific cutoff.

For the present analysis, left and right caudate volumes were summed to obtain a bilateral caudate measure, and the same procedure was applied to the putamen. The resulting bilateral volumes were then normalized by total intracranial volume to maintain consistency with the operational HD-ISS formulation, in which caudate and putamen volumes are expressed relative to intracranial volume.

The published age-specific cutoffs were then applied to the normalized

bilateral caudate and putamen measures. Because the present cohort was normative and did not include genetically defined Huntington's disease cases, this procedure was not used to assign a full HD-ISS stage, but rather to identify subjects meeting the Stage 1 imaging condition, here interpreted as neurodegeneration-positive (N+). A subject was therefore classified as N+ when either normalized putamen volume or normalized caudate volume fell below the corresponding age-dependent cutoff. The proportion of N+ classifications was then evaluated in the overall cohort and across datasets, to test whether the expected behaviour of the published thresholds was preserved in a heterogeneous multi-site normative sample. Possible sources of non-portability were subsequently investigated through statistical assessment of site effects and comparison of harmonization strategies.

3.2.2 Statistical assessment of site effects and evaluation of harmonisation strategies

This subsection investigated whether site-related variability affected the aggregated multi-site normative sample and evaluated harmonisation strategies aimed at improving the portability of cutoff-based pipelines across datasets. First, the presence of site effects was formally assessed. Different harmonisation methods were then applied and compared on normative structural MRI data.

Statistical assessment of site-related variability

Before applying harmonisation, site-related variability was formally assessed across the structural MRI features included in the dataset. This step served both to verify the presence of batch effects in the aggregated multi-site sample and to define a common statistical framework for the subsequent evaluation of harmonisation performance.

For each feature, site effects were tested using analysis of covariance (ANCOVA) implemented as a fixed-effects ordinary least squares model, with dataset as the categorical batch factor and age and sex as biological covariates to be preserved. The general model form was:

$$Y_{\text{feature}} = \beta_0 + \beta_1 \text{Age} + \beta_2 \text{Sex} + \sum_{k=1}^{K-1} \gamma_k \text{Dataset}_k + \varepsilon \tag{3.1}$$

where:

- Y_{feature} denotes the value of the structural feature for each subject;
- γ_k represents the effect of the acquisition dataset;
- ε is the residual error term.

In the software implementation, this model was specified as

$$\text{feature} \sim C(\text{dataset}) + \text{age} + \text{sex} \tag{3.2}$$

where $C(\text{dataset})$ indicates that dataset was treated as a categorical variable. ANCOVA tables were computed using Type II sums of squares (`typ=2`).

From each model, the F statistic, p-value, and effect size (η^2) associated with the dataset term were extracted to quantify the magnitude of site-related variability. Coefficients associated with age and sex were also monitored to verify preservation of biological relationships. The same procedure was repeated twice: first on the raw data, to assess the site effect before harmonisation, and then on the harmonised data, to evaluate the reduction or removal of the site effect. This implementation enabled direct comparison, for each feature, of the magnitude and statistical significance of site effects before and after harmonisation.

ComBat harmonization

ComBat harmonisation was implemented using the `neuroCombat.py` script [10]. The method was applied to the full set of structural MRI features in a single computational run, while estimating batch adjustment separately for each feature. Dataset was specified as the batch variable, whereas age and sex were included as biological covariates to be preserved.

Operationally, the procedure consisted of design matrix construction, feature-wise standardization, and empirical Bayes estimation of additive and multiplicative site effects. A standard parametric empirical Bayes ComBat model was adopted, with adjustment of both mean and variance and no reference batch.

Method 1 ComBat implementation settings

Require: full structural MRI feature set
Require: batch variable: dataset
Require: biological covariates: age, sex
1: implementation ← `neuroCombat.py` ▷ Fortin et al. (2018)
2: empirical Bayes ← `True`
3: parametric priors ← `True`
4: mean-only correction ← `False`
5: reference batch ← `None`

ComBat-GAM harmonization

ComBat-GAM harmonisation was implemented using `neuroHarmonize` package [14]. The method was applied to the full set of structural MRI features, with dataset specified as the batch variable and age and sex included as biological covariates to be preserved. In contrast to standard ComBat, age was modeled as a non-linear smooth term, allowing the harmonisation procedure to account for non-linear age-related effects across the lifespan.

The pipeline involved model estimation through `harmonizationLearn()`, followed by application of the learned harmonisation parameters through `harmonizationApply()`. A cubic spline basis was used for age, with empirical Bayes estimation enabled, mean and variance adjustment applied, and no reference batch.

Method 2 ComBat-GAM implementation settings

Require: full structural MRI feature set
Require: batch variable: dataset
Require: biological covariates: age, sex
1: implementation ← `neuroHarmonize` ▷ Pomponio et al. (2020)
2: model estimation ← `harmonizationLearn()`
3: model application ← `harmonizationApply()`
4: smooth terms ← `age`
5: spline basis ← `BSplines(df = 10, degree = 3)`
6: empirical Bayes ← `True`
7: mean-only correction ← `False`
8: reference batch ← `None`

CovBat harmonization

CovBat harmonisation was implemented using `covbat.py` [13]. The method was applied to the full set of structural MRI features, with dataset specified as the batch variable and age and sex included as biological covariates to be preserved.

The procedure first applied ComBat-based mean and variance correction, and then performed principal component analysis (PCA) on the standardized residuals in order to harmonise site-related differences in covariance structure. ComBat adjustment was subsequently applied to the principal component scores up to a predefined cumulative variance threshold. In the present implementation, automatic component selection was used, with harmonisation applied up to 95% of explained variance.

Method 3 CovBat implementation settings

Require: full structural MRI feature set
Require: batch variable: dataset
Require: biological covariates: age, sex
1: implementation ← `covbat.py` ▷ Chen et al. (2022)
2: numerical covariates ← `age`
3: cumulative variance threshold ← `0.95`
4: fixed number of PCs ← `0`
5: PCA on standardized residuals ← `True`
6: covariance harmonisation ← `True`

Comparative evaluation of harmonisation performance

Harmonisation performance was evaluated by reapplying the ANCOVA framework described above to the datasets obtained after ComBat, ComBat-GAM, and CovBat. This allowed site-related variability to be assessed using the same statistical criterion adopted for the raw data, thereby enabling direct comparison of residual batch effects across harmonisation methods.

For each structural feature, the F statistic, p-value, and effect size (η^2) associated with the dataset term were extracted from the post-harmonisation ANCOVA models. Coefficients associated with age and sex were also monitored to verify preservation of the biological relationships of interest.

In addition to ANCOVA, harmonisation performance was summarised using the Harmonization Effectiveness Score (HES), computed from the

pre- and post-harmonisation η^2 values of the dataset term. In the present workflow, HES was used as a complementary quantitative index of site-effect reduction, together with residual statistical significance and residual effect size. The three harmonisation methods were therefore compared on the basis of residual dataset effect, reduction in η^2, HES, and preservation of age- and sex-related associations.

3.2.3 Re-estimation of HD-ISS Stage 1 thresholds

This section describes the re-estimation of HD-ISS Stage 1 thresholds within the multi-site normative dataset. To this end, the sample was partitioned into training and held-out test subsets at the cohort level, age-dependent caudate and putamen thresholds were re-estimated on the training data, and the resulting cutoffs were subsequently applied to independent test cohorts.

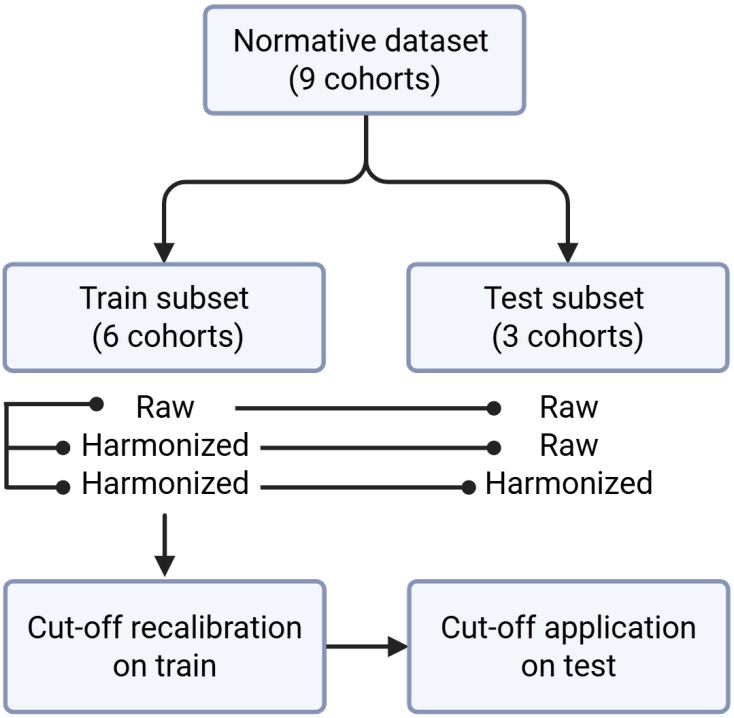


Figure 3.2. Methodological workflow for HD-ISS Stage 1 threshold re-estimation and held-out application.

The procedure was examined under three train-test configurations, including non-harmonized and harmonized settings, to assess the transferability of the recalibrated thresholds across previously unseen cohorts. The general rationale was aligned with the original HD-ISS framework, in which Stage 1 cutoffs are derived from normative data as age-dependent thresholds for striatal volumetric markers.

Train-test split

The normative sample was partitioned at the level of entire cohorts, rather than individual subjects, in order to preserve independence between threshold estimation and downstream application. This design allowed the recalibrated thresholds to be derived from one set of contributing cohorts and then evaluated on previously unseen held-out cohorts, thereby providing a more stringent assessment of cross-site transportability.

The training subset was defined to ensure adequate sample size and age coverage for stable estimation of age-dependent thresholds, whereas the test subset was kept fully independent for out-of-sample application. Cohort assignment to the training and held-out test subsets is reported in Table **??**. Additional details on the contributing cohorts, including MRI acquisition characteristics and demographic features, are provided in Tables 3.2 and **??** in the Materials section. Within this framework, the held-out subset also enabled evaluation on cohorts acquired under conditions that differed from much of the training sample.

Subset	**Cohorts**
Training	LA5c, NKI-RS, Cam-CAN, PREVENT-AD, SALD, OASIS-3
Held-out test	HCP-YoungAdult, HCP-Aging, DLBS

Table 3.5. Assignment of normative cohorts to the training and held-out test subsets.

Re-estimation of age-dependent thresholds

Age-dependent thresholds were re-estimated separately for caudate and putamen using quantile regression as implemented in the `rqpd` package in `R`, with the target quantile fixed at $\tau = 0.05$. The model was fitted using the pe-

nalized fixed-effects specification (`method = "pfe"`), and a constant penalty parameter ($\lambda = 20$) was retained across models as a fixed regularization setting.

For each striatal biomarker, the τ-th conditional quantile was modeled as a linear function of age:

$$Q_{Y_{ROI}}(\tau \mid \text{age}) = a_{ROI} + b_{ROI} \cdot \text{age}, \qquad \tau = 0.05 \tag{3.3}$$

where Y_{ROI} denotes the normalized volume of the biomarker of interest (either caudate or putamen), and a_{ROI} and b_{ROI} are the estimated intercept and age coefficient, respectively.

The age-dependent cutoff was then obtained directly from the fitted 5th-percentile regression line. Accordingly, for each ROI, the recalibrated threshold at a given age was defined as:

$$\text{cutoff}_{ROI}(\text{age}) = a_{ROI} + b_{ROI} \cdot \text{age} \tag{3.4}$$

Thus, the quantile regression model provided an age-specific estimate of the 5th-percentile normative boundary, which was subsequently used as the Stage 1 threshold for classification.

Application to the held-out test set under alternative harmonisation settings

The age-dependent thresholds estimated in the training subset were subsequently applied to the held-out test cohorts. For each subject, the expected cutoff was computed separately for caudate and putamen on the basis of age and the corresponding fitted coefficients. A subject was classified as Stage 1-like positive when at least one of the two striatal biomarkers fell below its age-specific threshold.

The procedure was examined under three train-test configurations:

- **raw-to-raw**, in which thresholds were estimated on non-harmonized training data and applied to non-harmonized test data;
- **harmonized-to-raw**, in which thresholds were estimated on harmonized training data and applied to non-harmonized test data;
- **harmonized-to-harmonized**, in which thresholds were estimated on harmonized training data and applied to harmonized test data.

These configurations were considered to separate the contribution of harmonisation at the threshold-estimation stage from its contribution at the test-application stage.

For the harmonized conditions, data were harmonized using the ComBat procedure implemented in the `neuroHarmonize` package, without specification of nonlinear smooth terms. The harmonization model was learned on the training subset and then applied to the held-out test data.

3.2.4 Hierarchical normative modelling approach

As an alternative to explicit data harmonisation, site-related variability was modelled directly within a hierarchical normative-modelling framework. To this end, Hierarchical Bayesian Regression (HBR) models [18] were implemented in PCNtoolkit [17] (https://github.com/predictive-clinical-neuroscience/PCNtoolkit) on the same multi-site normative sample described above. In order to ensure comparability with the recalibration analysis presented in Section 3.2.3, the training and held-out test subsets were kept identical.

The normative-modelling analysis was defined by the following settings:

- the same cohort-level train-test split adopted in Section 3.2.3;
- age, sex, and total intracranial volume as covariates;
- site as batch effect;
- caudate and putamen volumes as response variables;
- participant ID as subject identifier;
- two alternative likelihood specifications within the same HBR framework, namely a Gaussian model and a SHASH model;
- a nonlinear age effect in the location parameter, modelled using a B-spline basis with 5 knots and degree 3;
- posterior sampling with 4 chains, 1500 draws, 500 tuning iterations, and the `nutpie` NUTS sampler;

The SHASH likelihood was included as a more flexible alternative to the Gaussian specification for potentially non-Gaussian response distributions.

For each model, normative centiles and subject-level deviation scores were estimated in the held-out test cohorts. The downstream analysis focused on the 5th normative centile, and on the proportion of test subjects falling below this threshold for caudate and putamen.

3.2.5 Application of the recalibration pipeline to the iMarkHD sample

The final methodological step consisted of applying the recalibration pipeline to the independent *iMarkHD* structural MRI sample described in Subsection 3.1.4. In the original *iMarkHD* classification, group E comprised healthy controls, group B included participants in an early or transitional disease phase, and groups C and D represented more advanced disease stages [36]. For the present analysis, healthy controls from the *iMarkHD* subset (group E) were added to the control reference sample used for cutoff estimation, whereas participants from group B and the combined group C+D category were retained as the clinical test sample.

Before cutoff estimation, volumetric data were harmonized using ComBat, implemented through the Python `Harmonization` package. The harmonization model was learned on the control data with `harmonizationLearn` and transferred to the *iMarkHD* patient sample with `harmonizationApply`. No smooth term for age was included in the model. This step was intended to reduce unwanted site-related variability while preserving the biological covariates of interest [10].

Age-dependent cutoff functions for normalized caudate and putamen volumes were then re-estimated on the harmonized control data by means of quantile regression. This step followed the **HD-ISS!** (**HD-ISS!**) Stage 1 framework, in which striatal volumetric MRI markers are used as indicators of neurodegeneration and classification is based on age-adjusted thresholds derived from the lower tail of the normative distribution [4].

The recalibrated cutoff curves were finally applied to the harmonized *iMarkHD* patient groups. Classification was performed separately for caudate and putamen and then combined according to the original **HD-ISS!** Stage 1 rule, whereby a subject was considered positive for neurodegeneration when either marker fell below its corresponding age-adjusted cutoff [4].

4 Results

In this chapter, the results obtained from the evaluation of HD-ISS Stage 1 imaging cut-offs in a large multi-site normative MRI cohort are presented. The analysis first examines the performance of the original HD-ISS thresholds when directly applied to external normative data, and then quantifies the presence of site-related variability across structural MRI features.

The chapter then reports the comparative results of the harmonization strategies tested in this work, namely ComBat, CovBat, and ComBat-GAM, with performance assessed across the structural features included in the dataset. Building on these findings, the effects of harmonization on the portability of the original cut-offs are presented, followed by the results of their recalibration on external normative data under different training and test conditions.

In the final part of the chapter, the best-performing recalibration strategy is applied to independent clinical groups in order to evaluate its behaviour beyond the normative cohort. Lastly, the results obtained with hierarchical Bayesian normative modelling are presented as an alternative approach for accounting for site-related variability without explicit harmonization.

4.1 Performance of the Original HD-ISS Stage 1 Cut-offs in the Multi-Site Normative Cohort

To assess the portability of the original HD-ISS Stage 1 thresholds, the published age-dependent cut-offs were directly applied to the non-harmonized multi-site normative cohort. In the HD-ISS framework, Stage 1 is defined by biomarker evidence of pathogenesis captured by striatal volumetric MRI, with caudate and putamen identified as the relevant imaging landmarks. Stage 1 eligibility is determined according to an either-or rule, meaning that the condition is met when either caudate or putamen volume falls below the corresponding age-adjusted cut-off [4]. Accordingly, subjects meeting this criterion are referred to throughout this chapter as N+ subjects. In the original normative formulation, the nominal 5% reference applies directly to each imaging landmark considered separately, as the published thresholds were derived from the lower 5th percentile of the age-dependent normative distribution for caudate and putamen individually [4].

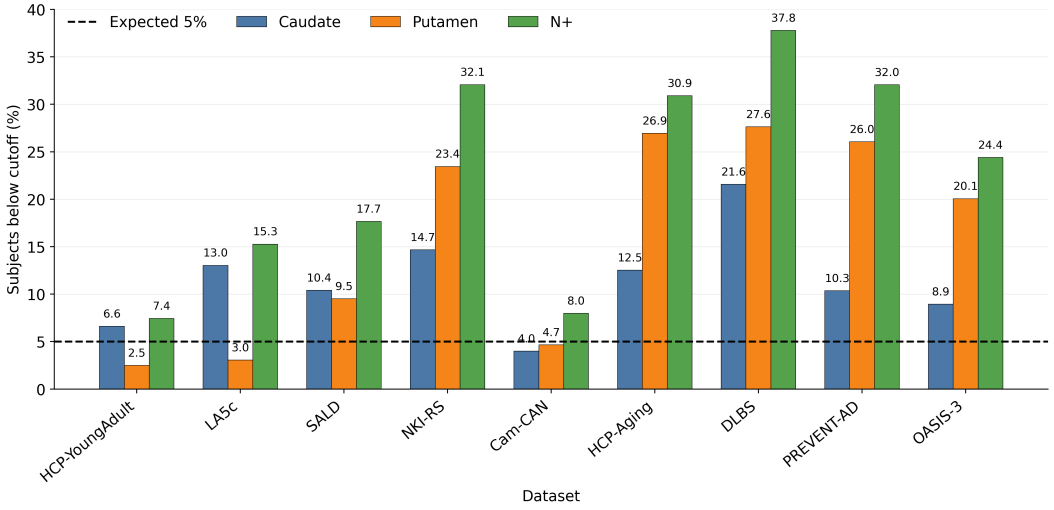


Figure 4.1. Classification rates obtained by applying the original HD-ISS Stage 1 cut-offs to the non-harmonized multi-site normative cohort. Percentages of subjects falling below the age-dependent cut-offs are shown separately for caudate and putamen, together with the percentage of subjects classified as N+ according to the Stage 1 either-or rule, across datasets and in the overall sample.

As shown in Figure 4.1 and Table 4.1, direct application of the original cut-offs to the external normative dataset yielded markedly non-uniform classification rates across the contributing datasets. The proportion of sub-

jects classified as N+ according to the Stage 1 either-or rule ranged from 7.44% in LA5c to 37.77% in HCP-YoungAdult. A similarly heterogeneous pattern was observed when the two imaging landmarks were considered separately: the percentage of subjects below the caudate cut-off ranged from 3.99% to 21.57%, whereas the proportion below the putamen cut-off ranged from 2.48% to 27.63%. Notably, landmark-specific rates were above the nominal 5% reference in most datasets, and the resulting N+ classification was also markedly inflated and highly variable across sites. Taken together, these findings indicate that the operational Stage 1 either-or rule was highly sensitive to between-dataset variability when transferred unchanged to the present external normative sample.

Across the pooled cohort, 13.06% of subjects fell below the caudate cut-off and 18.71% below the putamen cut-off, while 26.26% were classified as N+ according to the Stage 1 either-or rule. Thus, both individual landmark abnormality rates and the resulting N+ classification were substantially higher than would be expected under stable normative behaviour.

Table 4.1. Classification results obtained with the original HD-ISS Stage 1 cut-offs in the normative cohort. For each dataset, the table reports the sample size and the absolute and percentage number of subjects classified below the caudate cut-off, below the putamen cut-off, and classified as N+ because they met the Stage 1 either-or rule.

Dataset	N	n_{caud}	% caud	n_{put}	% put	$n_{\text{N+}}$	% N+
HCP-YoungAdult	1006	217	21.57	278	27.63	380	37.77
LA5c	121	8	6.61	3	2.48	9	7.44
SALD	491	64	13.03	15	3.05	75	15.27
NKI-RS	894	93	10.40	85	9.51	158	17.67
Cam-CAN	627	92	14.67	147	23.44	201	32.06
HCP-Aging	599	62	10.35	156	26.04	192	32.05
DLBS	414	37	8.94	83	20.05	101	24.40
PREVENT-AD	301	12	3.99	14	4.65	24	7.97
OASIS-3	631	79	12.52	170	26.94	195	30.90
Overall	5084	664	13.06	951	18.71	1335	26.26

Overall, these findings indicate that the original HD-ISS Stage 1 cut-offs did not show stable normative behaviour when applied unchanged to the present multi-site cohort. The inflation of landmark-specific abnormality

rates, together with the pronounced variability observed across datasets, provided an initial indication that direct transfer of the published thresholds to external data might be limited.

4.2 Site-Related Variability Across Structural MRI Features

Before harmonization, the raw structural MRI features displayed visible heterogeneity across the contributing datasets. Differences in both central tendency and dispersion were apparent across multiple measures, involving global volumetric variables as well as regional features. This pattern was not restricted to the striatal structures that are most directly relevant for the subsequent HD-ISS analyses, but extended to a broader set of structural measurements, indicating that between-dataset variability affected the pooled normative cohort more generally.

Figure 4.2 illustrates these differences for a representative subset of global and regional structural measures. Clear shifts between dataset-specific distributions were visible across multiple features, with only limited overlap between cohorts. Overall, the raw data exhibited marked between-dataset heterogeneity, which was subsequently assessed more formally through statistical analysis of site-related effects.

4.2.1 Statistical evidence of site effects in raw data

The descriptive heterogeneity observed in the raw feature distributions across datasets (Figure 4.2) was formally assessed using ANCOVA. Significant site-related effects were detected for all analysed structural MRI features (Table 4.2), indicating that dataset membership contributed systematically to the observed variability in the pooled normative cohort. The magnitude of the site effect, quantified by η^2, varied across features, ranging from more modest values for caudate volume ($\eta^2 = 0.012$) to larger effects for putamen volume ($\eta^2 = 0.042$), hippocampal volume ($\eta^2 = 0.106$), and especially temporal meta-ROI thickness ($\eta^2 = 0.410$).

Table 4.2 summarises the ANCOVA results obtained for the raw structural MRI features, reporting the statistical significance and effect size of the site

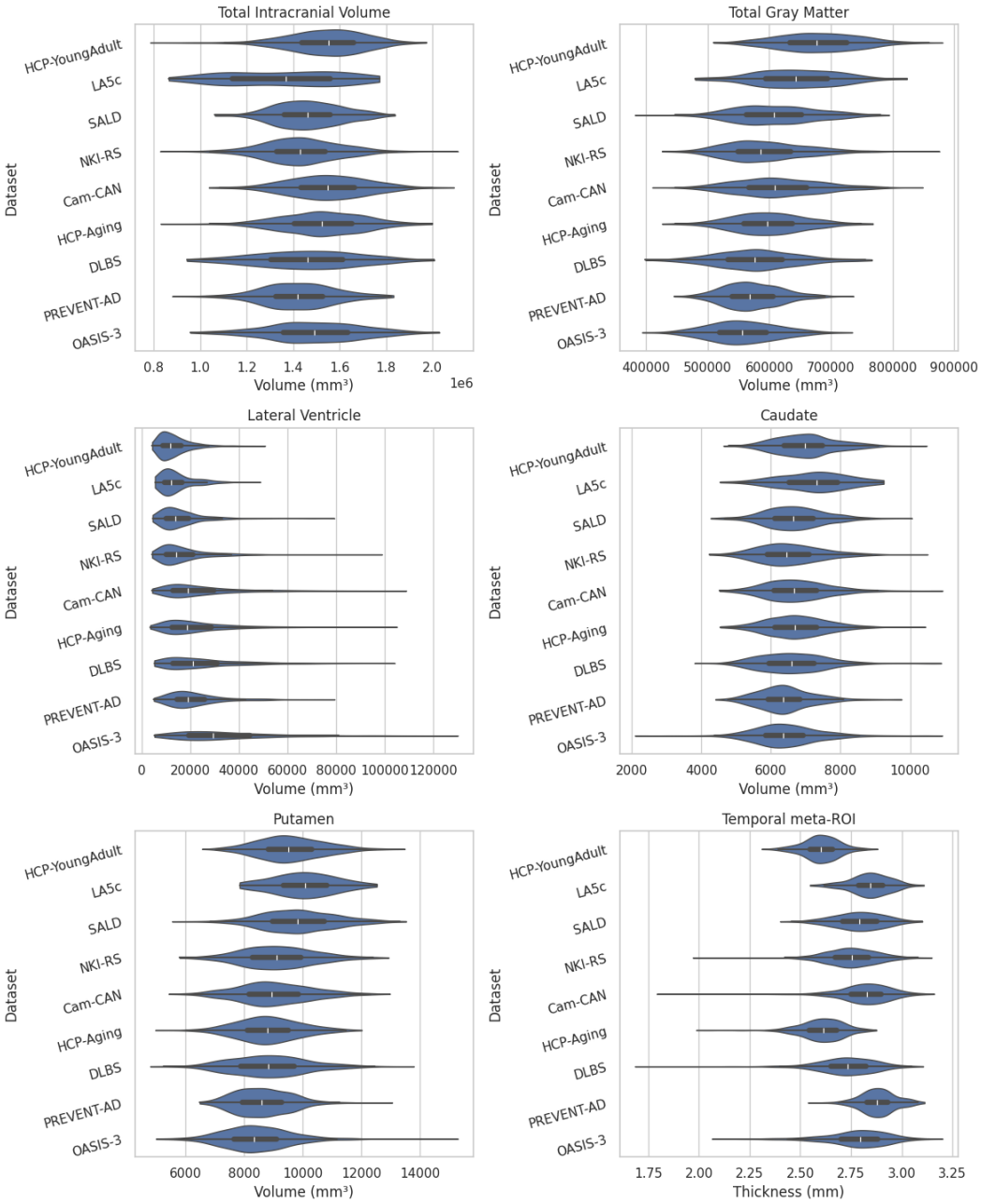


Figure 4.2. Distribution of a representative subset of raw structural MRI features across datasets before harmonization. Each panel shows the dataset-wise distribution of one feature included in the structural analysis. Visible differences in central tendency and spread were observed across several global and regional measures, indicating heterogeneity in the pooled normative cohort prior to harmonization.

term together with the contribution of age and sex.

Table 4.2. ANCOVA results for the raw structural MRI features. For each feature, the table reports the test statistic and p-value associated with the site effect, together with its effect size (η^2), and the regression coefficients and p-values for age and sex.

ROI	F	p_{site}	η^2_{site}	β_{age}	p_{age}	β_{sex}	p_{sex}
Total intracranial volume	50.99	7.11E-80	0.053	179.43	0.225	201111	<1E-300
Total gray matter volume	52.55	2.44E-82	0.043	-2144.1	<1E-300	60301.69	<1E-300
Lateral ventricle volume	30.12	1.95E-46	0.033	453.28	1.69E-294	6291.64	1.01E-78
Caudate volume	8.78	5.37E-12	0.012	-12.90	1.90E-50	541.29	1.16E-104
Putamen volume	42.21	6.91E-66	0.042	-39.79	<1E-300	846.12	1.98E-191
Hippocampus volume	99.32	3.52E-154	0.106	-27.14	2.08E-229	498.70	2.60E-102
Temporal meta-ROI thickness	529.78	<1E-300	0.410	-0.0037	5.87E-207	-0.0006	0.863

Note. F and p_{site} refer to the dataset term, while η^2_{site} quantifies the magnitude of the site effect. β_{age} and β_{sex} denote the regression coefficients for age and sex, respectively, with their associated p-values.

Among the two striatal measures directly relevant for the subsequent HD-ISS analyses, both caudate and putamen showed significant site effects, although with different effect sizes. In particular, the contribution of dataset membership was more pronounced for putamen than for caudate, consistent with the shifts already visible in the corresponding dataset-wise distributions in Figure 4.2.

Age was significantly associated with all structural measures except total intracranial volume ($p = 0.225$). The direction of the age coefficients was biologically plausible, with negative associations for total gray matter, caudate, putamen, hippocampus, and temporal meta-ROI thickness, and a positive association for lateral ventricle volume. Sex was significant for all volumetric measures included in the analysis, whereas no significant sex effect was observed for temporal meta-ROI thickness ($p = 0.863$). Notably, temporal meta-ROI differed from the other structural features: it represents a cortical thickness measure rather than a volumetric estimate, and therefore showed a partially distinct pattern with respect to the biological covariates.

Overall, these results confirmed the presence of widespread site-related effects in the raw structural MRI data, while still showing the expected contribution of biological covariates such as age and sex. Site-effect magnitude

was not uniform across structural MRI features (Figure 4.3), with the smallest effect observed for caudate volume and progressively larger values for putamen, hippocampus, and especially temporal meta-ROI thickness.

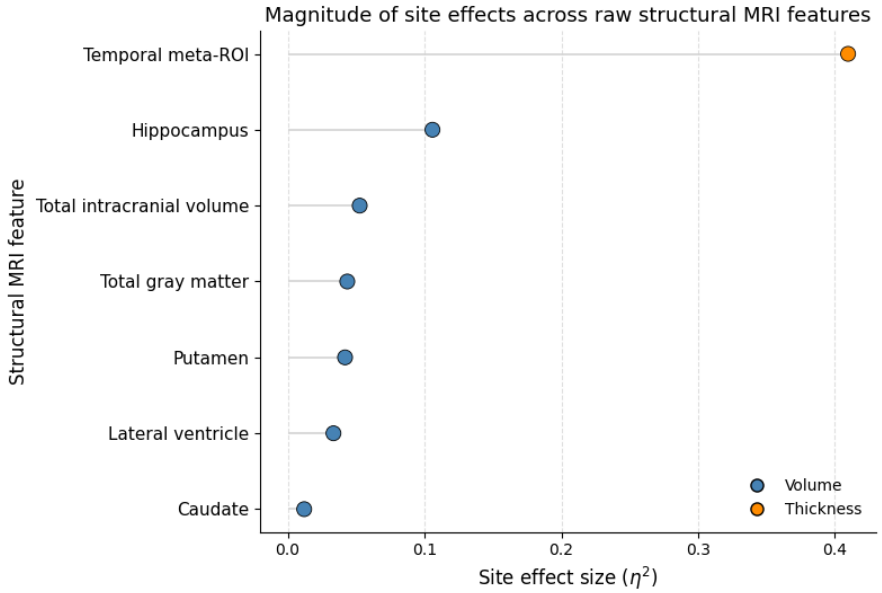


Figure 4.3. Magnitude of site effects across raw structural MRI features. The plot reports the η^2 associated with the dataset term for each feature in the ANCOVA model. Site-related variability was present across all analysed measures, with lower effect sizes for caudate volume and larger effects for putamen, hippocampus, and temporal meta-ROI thickness.

4.3 Comparative Evaluation of Harmonization Strategies Across Structural MRI Features

Given the widespread site-related variability observed in the raw structural MRI data, three harmonization strategies were evaluated: ComBat, CovBat, and ComBat-GAM. Their common objective was to attenuate site-related effects while retaining the contribution of biological covariates, particularly age and sex.

The post-harmonization results obtained with each method are presented below, followed by a comparative evaluation across structural features to identify the best-performing harmonization strategy.

4.3.1 Post-harmonization evaluation of ComBat

To assess the effect of ComBat harmonization on site-related variability, the ANCOVA analysis was repeated on the harmonized structural MRI features. In contrast with the results obtained on the raw data, no significant site effect was detected after ComBat for any of the analysed features. Consistently, the corresponding η^2 values were markedly reduced across the structural panel and approached zero in all cases. The full post-ComBat ANCOVA results are reported in Appendix Table 6.1.

This reduction was also evident in the features most relevant for the subsequent HD-ISS analyses. For caudate volume, the site-effect size decreased from $\eta^2 = 0.012$ before harmonization to $\eta^2 = 0.00001$ after ComBat, while for putamen volume it decreased from $\eta^2 = 0.042$ to $\eta^2 = 0.000007$. A similarly pronounced attenuation was observed for the other structural measures, including hippocampal volume and temporal meta-ROI thickness.

At the same time, the pattern of associations with the biological covariates remained largely preserved. Age remained non-significant for total intracranial volume, while it was still significant for all other features with biologically plausible directions of association. Sex also retained the same overall profile observed before harmonization, remaining significant for the volumetric features and non-significant for temporal meta-ROI thickness.

Overall, these results indicated that ComBat strongly attenuated site-related variability across structural MRI features while preserving the ex-

pected contribution of age and sex.

4.3.2 Post-harmonization evaluation of CovBat

To assess the effect of CovBat harmonization on site-related variability, the ANCOVA analysis was repeated on the CovBat-adjusted structural MRI features. Compared with the raw data, no significant site effect was detected after CovBat for any of the analysed features. Consistently, the corresponding η^2 values were markedly reduced across the entire structural panel and approached zero in all cases. The full post-CovBat ANCOVA results are reported in Appendix Table 6.3.

This attenuation was also evident in the features most relevant for the subsequent HD-ISS analyses. For caudate volume, the site-effect size decreased from $\eta^2 = 0.012$ before harmonization to $\eta^2 = 0.000019$ after CovBat, while for putamen volume it decreased from $\eta^2 = 0.042$ to $\eta^2 = 0.000003$. A similarly marked reduction was observed for the other structural measures, including hippocampal volume ($\eta^2 = 0.106$ to 0.000076) and temporal meta-ROI thickness ($\eta^2 = 0.410$ to 0.000136).

The pattern of associations with the biological covariates was also largely preserved after CovBat harmonization. Age remained non-significant for total intracranial volume ($p = 0.782$), while it was still significant for all other features with biologically plausible directions of association. In particular, negative age coefficients were observed for total gray matter, caudate, putamen, hippocampus, and temporal meta-ROI thickness, whereas lateral ventricular volume retained a positive association with age. Sex remained significant for all volumetric features, while no significant sex effect was observed for temporal meta-ROI thickness ($p = 0.414$).

Taken together, these results suggest that CovBat effectively mitigated site-related differences across structural MRI features without removing the biologically meaningful contributions of age and sex.

4.3.3 Post-harmonization evaluation of ComBat-GAM

To assess the effect of ComBat-GAM harmonization on site-related variability, the ANCOVA analysis was repeated on the ComBat-GAM-adjusted structural MRI features. Compared with the raw data, the magnitude of the

site effect was strongly reduced across the analysed features, with very small corresponding η^2 values in all cases. However, unlike the results observed after ComBat and CovBat, residual significant site effects were still detected for a subset of features (Appendix Table 6.2).

In particular, no significant site effect was detected after ComBat-GAM for total intracranial volume, total gray matter volume, or putamen volume. By contrast, residual site-related variability remained significant for lateral ventricular volume ($p = 4.82 \times 10^{-11}$), caudate volume ($p = 0.026$), hippocampal volume ($p = 1.00 \times 10^{-5}$), and temporal meta-ROI thickness ($p = 0.0096$). Although the associated effect sizes remained very small, these findings indicated that the site effect was not fully removed for all features.

Among the two striatal measures directly relevant for the subsequent HD-ISS analyses, putamen volume no longer showed a significant site effect after ComBat-GAM, whereas caudate volume retained a weak but significant residual association with dataset membership. This pattern contrasted with the complete attenuation observed after ComBat and CovBat.

The associations with the biological covariates remained largely preserved after ComBat-GAM harmonization. Age was still non-significant for total intracranial volume ($p = 0.565$), while it remained significant for all other features, with negative coefficients for total gray matter, caudate, putamen, hippocampus, and temporal meta-ROI thickness, and a positive coefficient for lateral ventricular volume. Sex also retained the same overall pattern observed before harmonization, remaining significant for all volumetric features and non-significant for temporal meta-ROI thickness ($p = 0.790$).

Overall, these results indicated that ComBat-GAM substantially reduced site-related variability, but did not completely eliminate residual site effects across all structural MRI features.

4.3.4 Comparative summary of harmonization performance

To compare the three harmonization strategies on a common basis, their performance was evaluated across the analysed structural MRI features using summary measures of residual site-related variability and the Harmonization Efficacy Score (HES). While all three approaches reduced the site effect relative to the raw data, their performance was not uniform across features and methods.

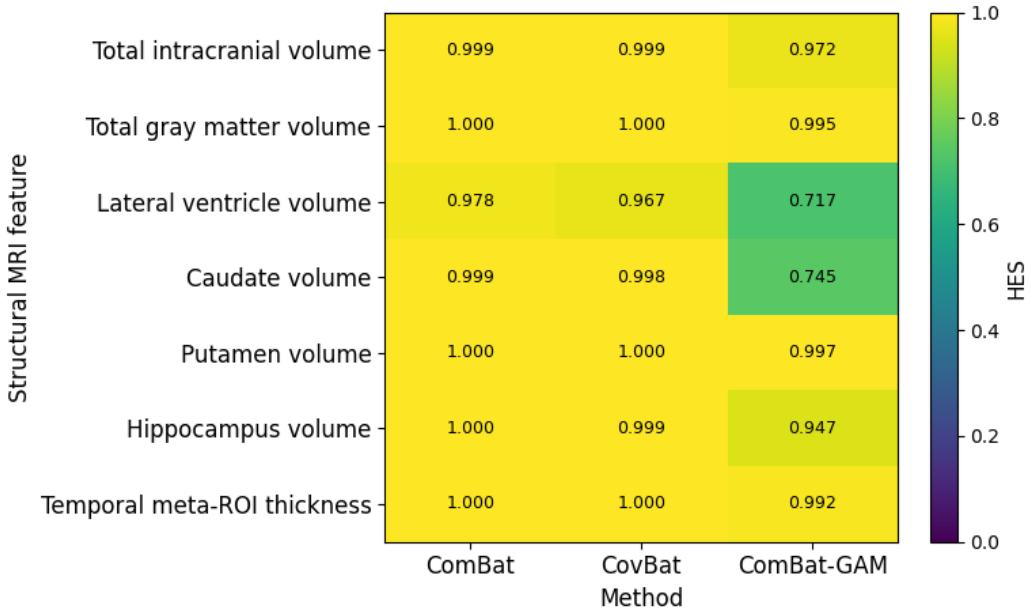


Figure 4.4. Comparative performance of the three harmonization strategies across structural MRI features. The heatmap reports the Harmonization Efficacy Score (HES) for each feature and method, with larger values indicating greater reduction of site-related variability relative to the raw data.

As shown in Figure 4.4, reduction of site-related variability was observed for all three methods, but with different patterns across the structural panel.

ComBat showed the most favorable overall profile, with consistently high harmonization efficacy across features. CovBat also produced a substantial attenuation of the site effect, although with a slightly less favorable overall pattern. By contrast, ComBat-GAM showed a less complete reduction of site-related variability, in agreement with the residual significant site effects observed in a subset of features in the post-harmonization ANCOVA.

Table 4.3. Summary comparison of harmonization performance across methods. For each method, the table reports summary statistics of post-harmonization site-effect size and Harmonization Efficacy Score (HES), together with the resulting overall ranking.

Method	Mean η^2_{site}	Median η^2_{site}	Mean HES	Median HES	Overall ranking
ComBat	0.000131	0.000047	0.9964	0.9995	1
CovBat	0.000204	0.000069	0.9946	0.9993	2
ComBat-GAM	0.003315	0.003051	0.9093	0.9721	3

Table 4.3 summarizes the overall comparative results and confirms the ranking suggested by the feature-wise analysis. On this basis, ComBat

was identified as the best-performing harmonization strategy among those tested and was therefore retained as the reference harmonized dataset for the subsequent analyses.

The post-ComBat distributions shown in Figure 4.5 provide a visual counterpart to the quantitative comparison, showing greater overlap across datasets than that observed in the raw data.

Taken together, these results indicated that ComBat achieved the most effective overall attenuation of site-related variability among the tested methods while preserving the expected structure of the biological covariates, and was therefore selected for the subsequent stages of the analysis.

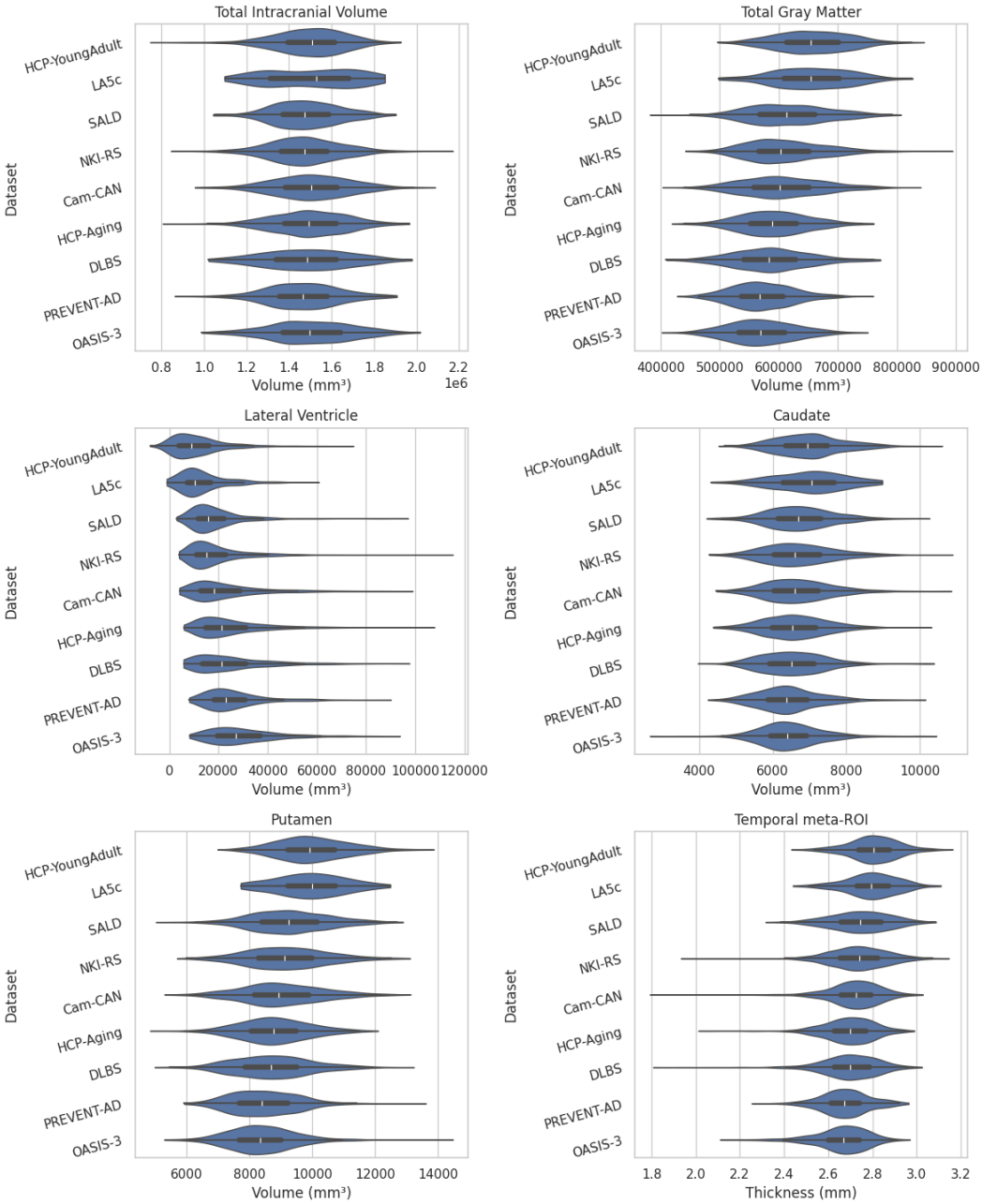


Figure 4.5. Distribution of structural MRI features across datasets after ComBat harmonization. The figure mirrors the raw-data visualization shown in Figure 4.2 and illustrates the improved alignment of the dataset-wise distributions after harmonization.

4.4 Effect of ComBat Harmonization on the Portability of the Original HD-ISS Stage 1 Cut-offs

After identifying ComBat as the best-performing harmonization strategy among those tested, the original HD-ISS Stage 1 age-dependent cut-offs were reapplied unchanged to the ComBat-harmonized normative cohort. The aim of this step was to assess whether removal of site-related variability alone was sufficient to restore the expected normative operating characteristics of the published Stage 1 thresholds [4].

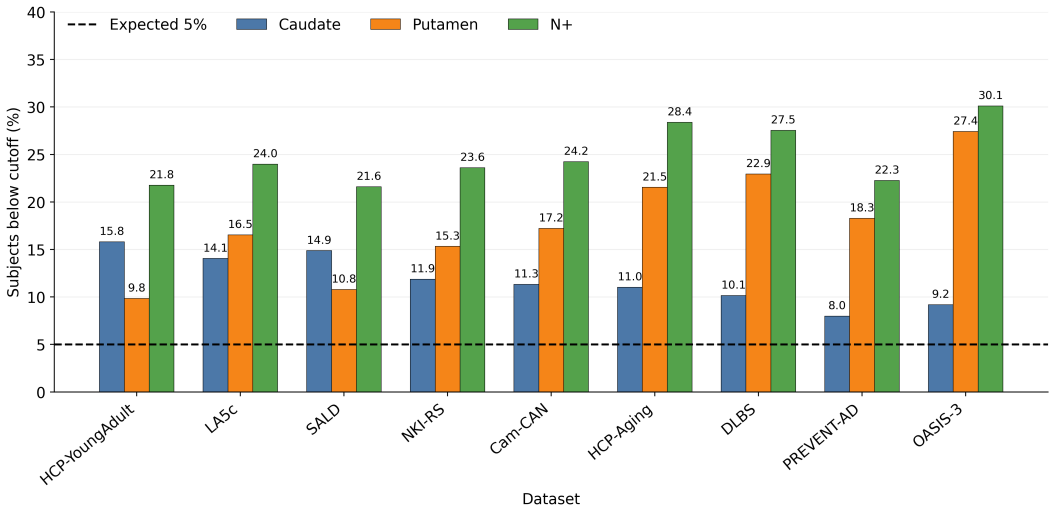


Figure 4.6. Classification rates obtained by applying the original HD-ISS Stage 1 cut-offs to the ComBat-harmonized normative cohort. Percentages of subjects falling below the age-dependent caudate and putamen cut-offs are shown separately, together with the percentage of subjects classified as N+ according to the HD-ISS Stage 1 either-or rule, across datasets. The dashed horizontal line marks the nominal 5% reference level for the individual imaging landmarks.

As shown in Figure 4.6 and Table 4.4, direct application of the original cut-offs to the harmonized data did not restore the expected normative behaviour of the published thresholds. The proportion of subjects classified as N+ according to the Stage 1 either-or rule remained markedly elevated in all datasets, ranging from 21.59% in SALD to 30.11% in OASIS-3. Similarly, the proportions falling below the caudate cut-off ranged from 7.97% to 15.81%, whereas those below the putamen cut-off ranged from 9.84% to 27.42%. In the overall harmonized sample, the corresponding proportions were 12.12% for caudate, 17.09% for putamen, and 24.74% for N+ subjects.

Table 4.4. Classification results obtained with the original HD-ISS Stage 1 cut-offs in the ComBat-harmonized normative cohort. For each dataset, the table reports the sample size and the absolute and percentage number of subjects classified below the caudate cut-off, below the putamen cut-off, and classified as N+ because they met the HD-ISS Stage 1 either-or rule.

Dataset	N	n_{caud}	% caudate	n_{put}	% putamen	$n_{\text{N+}}$	% N+
HCP-YoungAdult	1006	159	15.81	99	9.84	219	21.77
LA5c	121	17	14.05	20	16.53	29	23.97
SALD	491	73	14.87	53	10.79	106	21.59
NKI-RS	894	106	11.86	137	15.32	211	23.60
Cam-CAN	627	71	11.32	108	17.22	152	24.24
HCP-Aging	599	66	11.02	129	21.54	170	28.38
DLBS	414	42	10.14	95	22.95	114	27.54
PREVENT-AD	301	24	7.97	55	18.27	67	22.26
OASIS-3	631	58	9.19	173	27.42	190	30.11
Overall	5084	616	12.12	869	17.09	1258	24.74

Compared with the raw-data analysis presented in Section 4.1, harmonization slightly reduced the overall N+ classification rate and attenuated part of the between-dataset spread, but it did not re-centre the classification output around the intended normative target for the individual imaging landmarks. Instead, the resulting proportions remained systematically elevated across the full harmonized cohort.

Overall, these findings indicate that harmonization alone, even when based on the best-performing method identified in the previous section, was not sufficient to recover the intended operating characteristics of the original HD-ISS Stage 1 cut-offs on this external normative dataset. In particular, even after ComBat harmonization, application of the published Stage 1 either-or rule continued to classify a substantially larger-than-expected proportion of normative subjects as N+, thereby motivating the subsequent recalibration of the thresholds on external normative data.

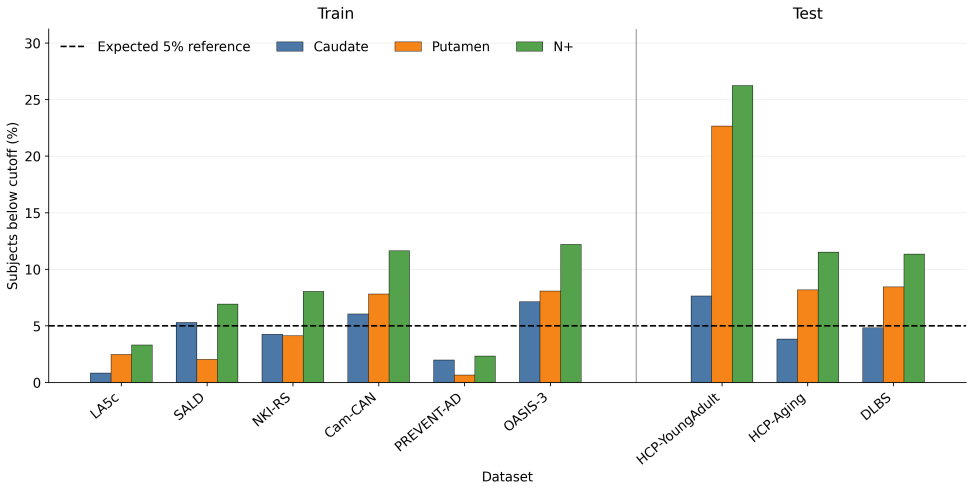
4.5 External recalibration of the HD-ISS Stage 1 cut-offs

To further assess whether the portability of the HD-ISS Stage 1 thresholds could be improved through external recalibration, the normative cohort was divided into independent training and test subsets. New age-dependent cut-offs were estimated on the training data and subsequently applied to the held-out test data under three different configurations: raw-to-raw, harmonized-to-raw, and harmonized-to-harmonized. In the original HD-ISS framework, Stage 1 entry is based on age-dependent control-derived cut-offs for caudate and putamen considered separately, while the stage condition is then evaluated according to an either-or rule across the two landmarks [4]. From this section onward, datasets are displayed according to this train/test partition, with training datasets shown first and test datasets afterwards, rather than by mean age as in the previous analyses.

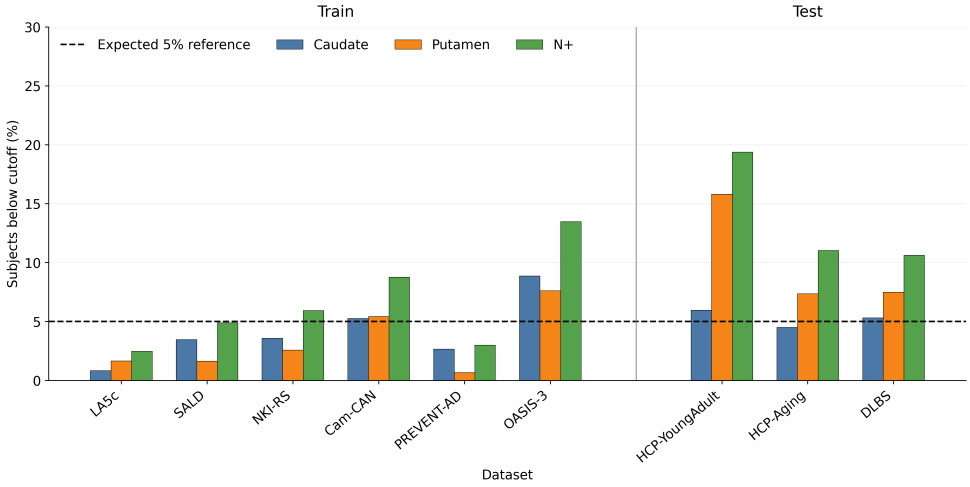
Figure 4.7 summarizes the classification profiles observed in the three scenarios, whereas Table 4.5 reports the overall train/test summary. For conciseness, only the aggregate comparison is presented in the main text, while the full dataset-specific results for the raw-to-raw, harmonized-to-raw, and harmonized-to-harmonized scenarios are reported in Appendix Tables 6.4, 6.5, and 6.6, respectively. Figure 4.7 displays, for each configuration, the percentages below the caudate and putamen cut-offs together with the resulting N+ classification; the 5% reference is shown as a visual benchmark for the individual imaging landmarks.

As shown in Figure 4.7 and Table 4.5, external recalibration improved the behaviour of the Stage 1 thresholds across all three configurations, although to different extents.

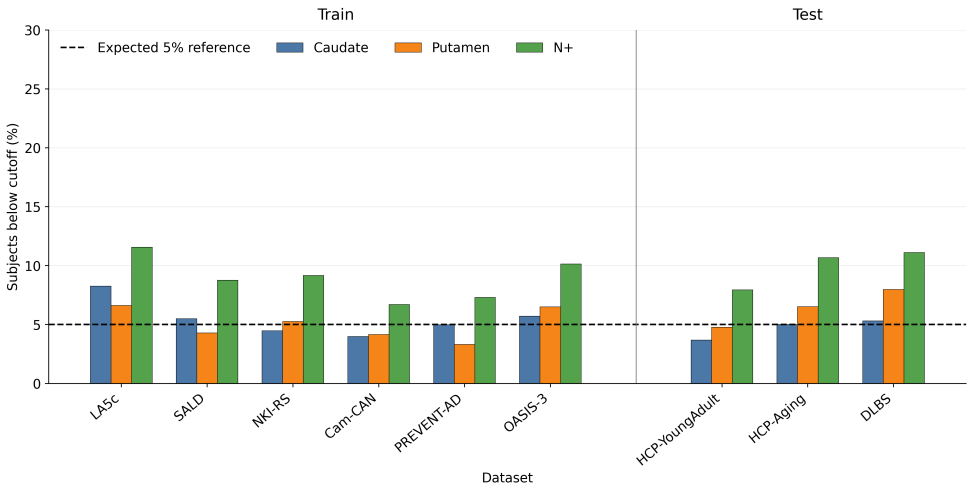
In the raw-to-raw setting, recalibration on external normative data already improved the behaviour of the thresholds relative to the direct application of the original HD-ISS cut-offs. However, test-set classification rates remained clearly inflated, with an overall N+ proportion of 18.82%, compared with 8.71% in the training data. This residual inflation was driven primarily by the putamen-based classification, which remained markedly elevated in the test set (15.45%), and was most evident in HCP-YoungAdult, where the N+ rate reached 26.24%. Thus, although retraining alone im-



(a) Raw-to-raw



(b) Harmonized-to-raw



(c) Harmonized-to-harmonized

Figure 4.7. Classification rates obtained after external recalibration of the HD-ISS Stage 1 cut-offs under three train/test configurations.

Table 4.5. Overall train/test summary of the recalibrated HD-ISS Stage 1 thresholds across the three evaluated configurations. For each scenario, the table reports the overall percentage of N+ subjects in the training set, the overall test-set percentages below the caudate and putamen cut-offs, the overall percentage of subjects classified as N+ in the test set, and the range of test-set N+ rates across the three held-out datasets. Full dataset-specific results are reported in Appendix Tables 6.4–6.6.

Scenario	Train N+ (%)	Test caudate (%)	Test putamen (%)	Test N+ (%)	Test N+ range (%)
Raw-to-raw	8.71	5.94	15.45	18.82	11.35–26.24
Harmonized-to-raw	7.47	5.40	11.59	15.11	10.63–19.38
Harmonized-to-harmonized	8.71	4.41	5.94	9.41	7.95–11.11

proved transportability, substantial out-of-sample instability persisted.

A partial improvement was observed when cut-offs were estimated on harmonized training data but applied to non-harmonized test data. Under this configuration, the overall N+ rate in the test set decreased to 15.11%, and the spread across the three test datasets narrowed to 10.63%–19.38%. The putamen contribution was also attenuated relative to the raw-to-raw case, but it remained the main driver of inflated classification in the held-out data (11.59% versus 5.40% for caudate). These findings suggest that harmonization at the estimation stage was beneficial, but that applying the recalibrated thresholds to test data still represented in the raw feature space limited the achievable gain.

The most coherent results were obtained in the harmonized-to-harmonized configuration. In this case, the overall N+ proportion in the independent test set was reduced to 9.41%, while the range across the three held-out datasets narrowed further to 7.95%–11.11%. Importantly, the discrepancy between training and test became minimal in this scenario (8.71% versus 9.41%), indicating much better alignment between threshold estimation and out-of-sample application. A similar stabilization was observed for the separate caudate and putamen classifications, whose overall test-set rates were 4.41% and 5.94%, respectively, that is, much closer to the nominal landmark-specific reference than in the other two configurations. Notably, HCP-YoungAdult showed the strongest reduction in N+ rate across scenarios, from 26.24% in

the raw-to-raw setting to 19.38% in the harmonized-to-raw setting and 7.95% in the harmonized-to-harmonized setting.

To facilitate interpretation of the recalibration step, the estimated parameters of the age-dependent Stage 1 cut-offs are reported separately for the original HD-ISS model (Table 4.6) and for the two recalibrated models estimated on the raw and harmonized training sets (Table 4.7).

Table 4.6. Original age-dependent HD-ISS Stage 1 cut-off parameters, modeled as cutoff$(age) = a + b \cdot age$, where a is the intercept and b is the slope.

	Original HD-ISS	
ROI	Intercept (a)	Slope (b)
Caudate	0.004558	-1.3×10^{-5}
Putamen	0.006364	-1.8×10^{-5}

Table 4.7. Recalibrated age-dependent HD-ISS Stage 1 cut-off parameters estimated on the raw and harmonized training sets. Cut-offs were modeled as cutoff$(age) = a+b\cdot age$, where a is the intercept and b is the slope. The parameters estimated on the harmonized training set were used in both the harmonized-to-raw and harmonized-to-harmonized configurations.

	Retrained on raw train		Retrained on harm train	
ROI	Intercept (a)	Slope (b)	Intercept (a)	Slope (b)
Caudate	0.004201008	-9.934876×10^{-6}	0.004062410	-7.424964×10^{-6}
Putamen	0.006598793	-2.867884×10^{-5}	0.006385557	-2.598803×10^{-5}

Relative to the original HD-ISS model, both recalibrated solutions yielded lower caudate intercepts and flatter age-dependent caudate slopes. For putamen, the raw-train recalibration produced both a higher intercept and a steeper age-related decline than the original model, whereas the harmonized-train recalibration preserved an intercept close to the published value while still yielding a steeper slope. These differences indicate that external recalibration modified not only the absolute level of the cut-offs but also their age dependence, consistent with the improved out-of-sample behaviour observed for the harmonized training pipeline.

Taken together, these results show a progressive improvement across the three evaluated configurations. Three patterns emerged. First, recalibration in the raw-to-raw setting already reduced misclassification rates relative to

direct transfer of the original thresholds. Second, a further improvement was observed when cut-offs were estimated on harmonized training data. Finally, the best results were obtained when both training and test data were represented in the same harmonized space. This last configuration provided the closest approximation to the intended HD-ISS Stage 1 normative behaviour, which is based on age-dependent control-derived cut-offs for caudate and putamen, with N+ subjects defined according to the either-or rule across the two landmarks [4]. On this basis, the harmonized-to-harmonized configuration was retained for the subsequent analyses as the most appropriate implementation of externally recalibrated Stage 1 cut-offs.

4.6 Normative Modelling Without Explicit Harmonization

As an alternative to explicit data harmonization, hierarchical normative modelling was evaluated to determine whether scanner- and site-related variability could be accounted for directly at the modelling stage. Two Hierarchical Bayesian Regression (HBR) formulations were considered, namely a Gaussian model and a SHASH model. In both cases, the models were trained on the same non-harmonized training subset used in the previous analyses, including the LA5c, NKI-RS, Cam-CAN, PREVENT-AD, SALD, and OASIS-3 datasets, and were then tested on the same independent external sites, namely HCP-YoungAdult, HCP-Aging, and DLBS. This design ensured direct comparability with the harmonization-based and recalibration-based analyses reported in the previous sections.

Within the HD-ISS framework, Stage 1 is defined by MRI-based evidence of striatal neurodegeneration, operationalized through age-dependent volumetric landmarks for the caudate and putamen. These landmark thresholds were originally derived from the lower 5% tail of the control distribution, and Stage 1 eligibility is reached when either the caudate or the putamen falls below its corresponding age-dependent cut-off according to the either-or rule [4]. Accordingly, in the present analysis, model performance on the external healthy test cohorts was assessed by quantifying the proportion of subjects falling below the model-derived 5th-centile threshold for each ROI separately, together with the resulting proportion of N+ subjects under the Stage 1 either-or rule.

4.6.1 External-site calibration in healthy test cohorts

The first objective was to assess whether the normative models preserved adequate calibration when transferred to external healthy datasets acquired at unseen sites. A well-calibrated model should classify approximately 5% of healthy subjects as abnormal for each ROI separately, consistently with the tail probability used in the definition of the HD-ISS Stage 1 imaging landmarks. By contrast, the N+ rate is reported here as an operational analogue of Stage 1 classification, but it is not expected to coincide exactly with 5%,

because it combines two landmark-specific thresholds under the either-or rule [4]. Deviations from these expected patterns therefore provide a direct indication of residual site-related distortion in the normative predictions.

For the Gaussian HBR model, the proportion of subjects falling below the 5th-percentile threshold remained markedly heterogeneous across external sites. Under the Stage 1 either-or rule, the percentage of subjects classified as N+ was 10.87% in DLBS, 10.35% in HCP-Aging, and 27.14% in HCP-YoungAdult. A similar pattern was observed at the ROI level. For the caudate, the observed proportions were 4.35% in DLBS, 2.67% in HCP-Aging, and 9.84% in HCP-YoungAdult, whereas for the putamen they were 8.21%, 7.85%, and 21.87%, respectively. Overall, these findings indicate that the Gaussian model only partially restored the expected external calibration, with the largest deviation observed in the youngest test cohort and with a more pronounced inflation for the putamen than for the caudate.

The corresponding quantitative results for the Gaussian and SHASH formulations are summarised in Tables 4.8 and 4.9, respectively, while Figure 4.8 provides a direct visual comparison of abnormal classification rates across the three independent healthy test cohorts.

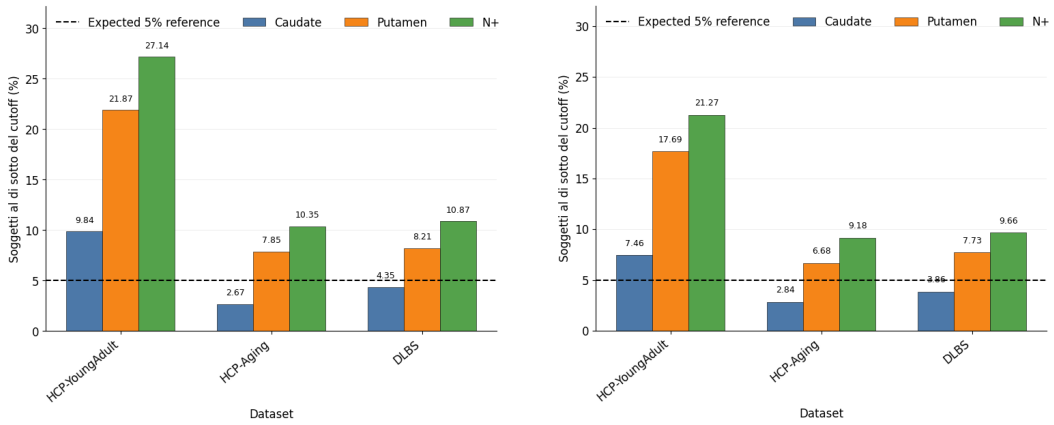
Table 4.8. External-site calibration of the HBR Gaussian model in the healthy test cohorts. For each dataset, the table reports the proportion of subjects classified below the model-derived 5th-percentile threshold for the caudate and putamen separately, together with the resulting N+ classification according to the either-or rule.

Dataset	N	Caudate		Putamen		N+	
		n	%	n	%	n	%
HCP-YoungAdult	1006	99	9.84	220	21.87	273	27.14
HCP-Aging	599	16	2.67	47	7.85	62	10.35
DLBS	414	18	4.35	34	8.21	45	10.87

Taken together, the Gaussian results indicate that hierarchical normative modelling can mitigate part of the cross-site variability, but does not fully eliminate residual site dependence in normative inference when applied without an explicit harmonization step. In particular, the strongest departure from the expected landmark-specific calibration was observed in HCP-YoungAdult, especially for the putamen and, consequently, for the resulting N+ classification.

Table 4.9. External-site calibration of the HBR SHASH model in the healthy test cohorts. For each dataset, the table reports the proportion of subjects classified below the model-derived 5th-percentile threshold for the caudate and putamen separately, together with the resulting N+ classification according to the either-or rule.

Dataset	N	Caudate		Putamen		N+	
		n	%	n	%	n	%
HCP-YoungAdult	1006	75	7.46	178	17.69	214	21.27
HCP-Aging	599	17	2.84	40	6.68	55	9.18
DLBS	414	16	3.86	32	7.73	40	9.66



(a) HBR Gaussian **(b)** HBR SHASH

Figure 4.8. External-site calibration of the HBR Gaussian and HBR SHASH models in the healthy test cohorts. For each model, the grouped bar plot reports the percentage of subjects classified below the model-derived 5th-percentile threshold for caudate and putamen, together with the percentage of N+ subjects according to the either-or rule, in HCP-YoungAdult, HCP-Aging, and DLBS. The horizontal reference line indicates the nominal 5% landmark-specific reference and is shown for visual comparison.

4.6.2 Visual inspection of normative centiles across external sites

To complement the percentage-based analysis, the age-dependent normative centiles predicted by the models were visually inspected in the external test cohorts. This representation makes it possible to assess whether subjects from unseen datasets remain approximately aligned with the expected normative bands across the age range, or whether systematic offsets persist at the level of individual sites. In the present context, these plots were used as qualitative support for the calibration results rather than as the primary summary measure.

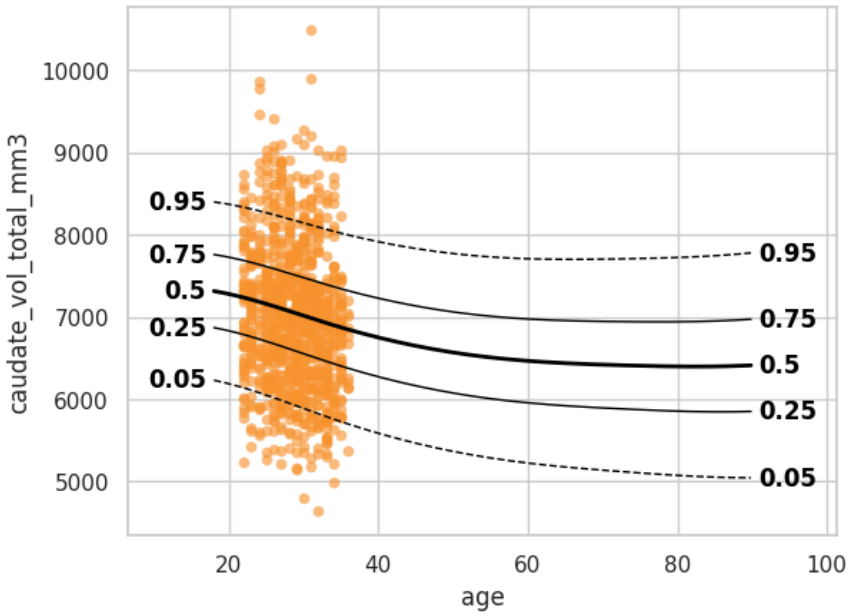
For clarity, only the centile plots obtained on the HCP-YoungAdult test cohort are shown in the main text, as this dataset displayed the largest deviation from the expected calibration pattern and therefore provided the most informative visual comparison between the Gaussian and SHASH formulations. The corresponding plots for HCP-Aging and DLBS are reported in Appendix Figures 6.10 and 6.11.

Visual inspection of the centile curves suggested that the HBR models were able to capture the expected age-related decline in striatal volumes, but that some residual dataset-dependent displacement remained across the external sites. For this reason, the centile plots were interpreted as supportive evidence of partial, rather than complete, correction of batch effects by the normative models.

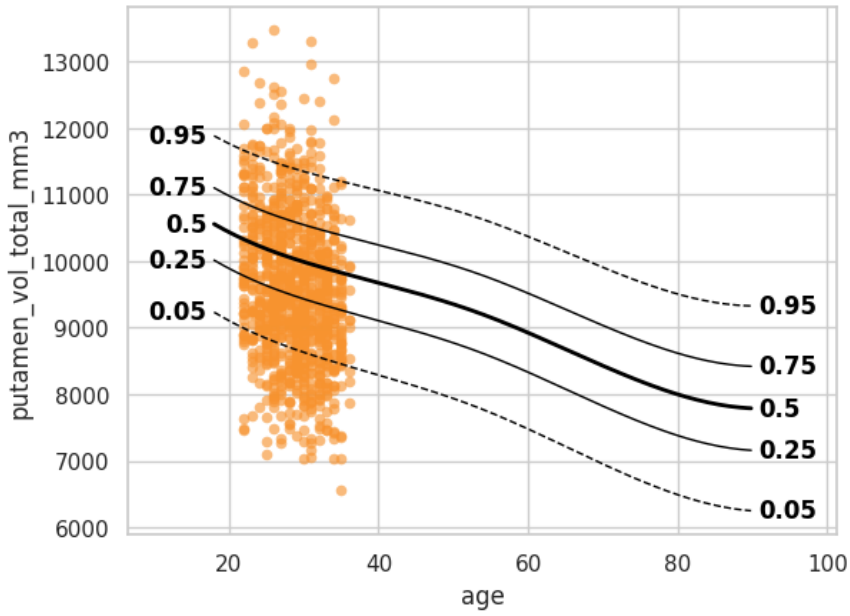
4.6.3 Comparison between Gaussian and SHASH formulations

A direct comparison between the Gaussian and SHASH formulations was performed to determine which model achieved calibration closer to the expected landmark-specific reference across the external healthy cohorts. In this setting, the preferred formulation should not simply provide an adequate fit to the training data, but should also produce a stable and portable normative reference when transferred to unseen sites. The comparison was therefore based primarily on external-site calibration, separately for the caudate, putamen, and N+ classification according to the either-or rule.

For the Gaussian model, the pooled proportions across the three external

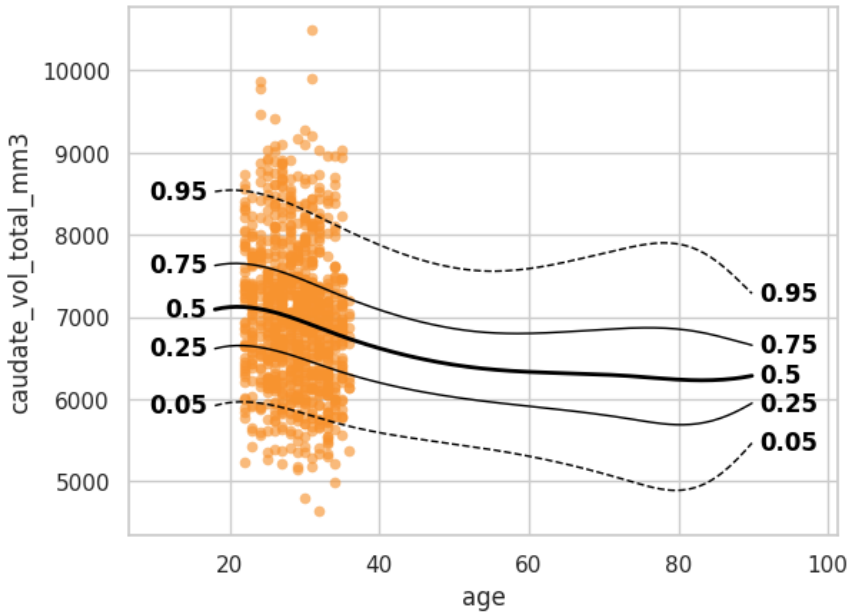


(a) Caudate

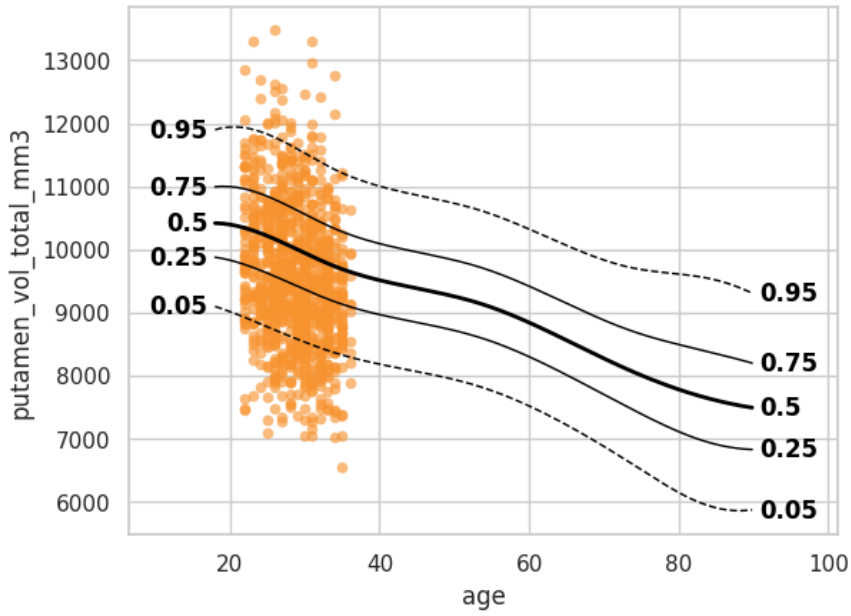


(b) Putamen

Figure 4.9. Representative age-dependent normative centile plots for the HCP-YoungAdult external test cohort obtained with the HBR Gaussian model, shown separately for caudate and putamen. HCP-YoungAdult is displayed in the main text because it showed the largest deviation from the expected calibration profile.



(a) Caudate



(b) Putamen

Figure 4.10. Representative age-dependent normative centile plots for the HCP-YoungAdult external test cohort obtained with the HBR SHASH model, shown separately for caudate and putamen. HCP-YoungAdult is displayed in the main text because it showed the largest deviation from the expected calibration profile.

test datasets were 6.59% for the caudate, 14.91% for the putamen, and 18.82% for N+ subjects according to the either-or rule, confirming a substantial inflation of abnormal classifications at the aggregate level. These pooled results further support the conclusion that the Gaussian HBR formulation was not sufficient to restore the expected landmark-specific calibration in healthy external cohorts, particularly for the putamen and, consequently, for the resulting N+ classification.

Overall, the SHASH formulation showed a more favorable external calibration profile than the Gaussian HBR model and therefore represented the better-performing normative modelling variant in the present study. However, despite this improvement, normative modelling without explicit harmonization remained less effective than the best harmonization-based recalibration pipeline. These results suggest that, in this specific multi-site application, HBR SHASH provided only a partial correction of batch effects and did not match the external stability achieved by the harmonized-to-harmonized recalibration approach.

Table 4.10. Summary comparison of the normative modelling strategies and the reference harmonization-based recalibration pipeline. Pooled percentages were computed across the three external healthy test cohorts.

Method	Caudate (%)	Putamen (%)	N+ (%)
HBR Gaussian	6.59	14.91	18.82
HBR SHASH	5.35	12.38	15.30
Best harmonization-based recalibration pipeline	4.41	5.94	9.41

Taken together, these findings indicate that hierarchical normative modelling can reduce the impact of batch effects on striatal normative inference, but only partially, and therefore does not replace the benefit of explicit harmonization in this specific multi-site application.

4.7 Application of the HD-ISS Stage 1 framework to the iMarkHD sample

This section reports the application of the HD-ISS Stage 1 framework to the *iMarkHD* sample under two conditions. First, the original published thresholds were applied directly to the normalized caudate and putamen volumes. Second, the recalibrated pipeline derived from the external normative analyses was applied after harmonization and local cut-off estimation. The resulting classifications are then compared at both the group and subject levels, and the corresponding age-dependent cut-off parameters are reported to characterize the shift introduced by recalibration.

4.7.1 Application of the original HD-ISS thresholds to the iMarkHD sample

The published HD-ISS Stage 1 thresholds [4] were first applied directly to the *iMarkHD* sample using the normalized caudate and putamen volumes. The resulting classification is summarized in Table 4.11.

A progressive increase in the proportion of subjects classified below the age-adjusted cut-offs was observed across the three clinical groups. Among healthy controls (group E), no subjects were below the caudate cut-off, whereas 2 out of 34 subjects (5.9%) fell below the putamen cut-off, yielding an overall N+ rate of 5.9% according to the HD-ISS either-or rule. In the early neurodegeneration group (group B), 6 out of 21 subjects (28.6%) were below the caudate cut-off and 9 out of 21 (42.9%) were below the putamen cut-off, with 11 subjects (52.4%) classified as N+. In the neurodegeneration group (group C+D), the corresponding proportions increased to 26 out of 37 subjects (70.3%) for the caudate, 25 out of 37 (67.6%) for the putamen, and 29 out of 37 (78.4%) classified as N+.

Overall, direct application of the original thresholds identified an increasing burden of striatal abnormality from controls to the two clinical groups, with the highest proportion of N+ classifications observed in subjects with more advanced neurodegeneration.

Table 4.11. Direct application of the original HD-ISS Stage 1 thresholds to the *iMarkHD* sample. For each group, the table reports the number and percentage of subjects below the age-adjusted caudate and putamen cut-offs, together with the resulting N+ classification according to the either-or rule.

Group	*N*	Caudate		Putamen		N+ (either-or rule)	
		n	%	*n*	%	*n*	%
Controls (E)	34	0	0.0	2	5.9	2	5.9
Early neurodegeneration (B)	21	6	28.6	9	42.9	11	52.4
Neurodegeneration (C+D)	37	26	70.3	25	67.6	29	78.4

4.7.2 Application of the recalibrated pipeline to the iMarkHD sample

The recalibrated pipeline was subsequently applied to the *iMarkHD* sample after harmonization and local cut-off estimation, as described in the Methods section. Because healthy controls from the *iMarkHD* subset were incorporated into the harmonized reference sample used for cut-off estimation, the quantitative results of the recalibrated application are reported here only for the two patient groups.

As summarized in Table 4.12, the recalibrated pipeline yielded higher proportions of subjects below the age-adjusted striatal cut-offs in both clinical groups. In the early neurodegeneration group (group B), 9 out of 21 subjects (42.9%) were below the caudate cut-off and 11 out of 21 (52.4%) were below the putamen cut-off, with 11 subjects (52.4%) classified as N+ according to the either-or rule. In the neurodegeneration group (group C+D), 30 out of 37 subjects (81.1%) were below the caudate cut-off, 31 out of 37 (83.7%) were below the putamen cut-off, and 33 out of 37 (89.2%) were classified as N+.

Relative to the direct application of the published thresholds, the recalibrated pipeline was associated with an upward shift in the proportion of subjects identified below the regional cut-offs, particularly in the group with more advanced neurodegeneration. This effect was especially evident for the putamen and for the combined N+ classification in group C+D, indicating a stronger separation between the two patient groups after recalibration.

Figures 4.11 and 4.12 provide a direct visual comparison between the two classification scenarios for the caudate and putamen, respectively. In each figure, the upper panel shows the application of the original published HD-ISS Stage 1 cut-off curve, whereas the lower panel shows the recalibrated

Table 4.12. Application of the recalibrated pipeline to the *iMarkHD* patient groups. For each group, the table reports the number and percentage of subjects below the age-adjusted caudate and putamen cut-offs, together with the resulting N+ classification according to the either-or rule.

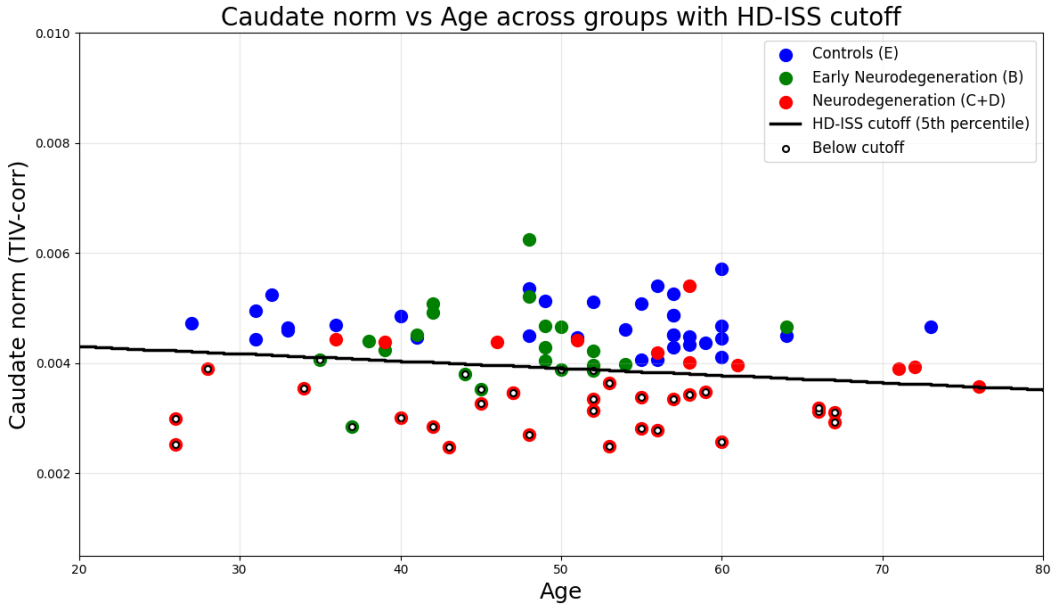
Group	*N*	Caudate		Putamen		N+ (either-or rule)	
		n	%	*n*	%	*n*	%
Early neurodegeneration (B)	21	9	42.9	11	52.4	11	52.4
Neurodegeneration (C+D)	37	30	81.1	31	83.7	33	89.2

cut-off curve after harmonization and local cut-off estimation. This representation complements the group-level summaries reported in Tables 4.11 and 4.12 by illustrating the shift in subject distribution with respect to the classification boundaries after recalibration.

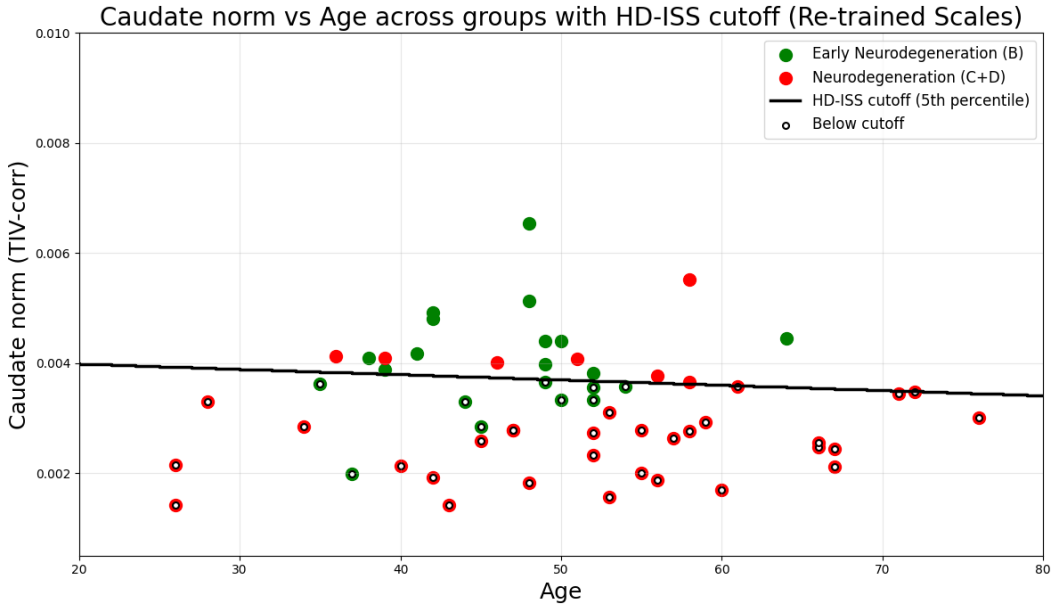
Visual inspection of Figures 4.11 and 4.12 showed that the recalibrated cut-off curves shifted the classification boundaries in a direction that increased the number of patient observations falling below the age-adjusted limits, particularly in the more advanced clinical group. In group B, this effect was visible at the regional level, with the proportion of subjects below cut-off increasing from 28.6% to 42.9% for the caudate and from 42.9% to 52.4% for the putamen, although the overall N+ rate remained unchanged at 52.4%. In contrast, in group C+D the effect extended to the combined classification, with the proportion of subjects below cut-off increasing from 70.3% to 81.1% for the caudate and from 67.6% to 83.7% for the putamen, while the overall N+ rate increased from 78.4% to 89.2%. Taken together, these patterns indicate that recalibration mainly improved the identification of striatal abnormality in the clinically more advanced patients, with the largest gain observed for putaminal involvement and for the final N+ classification in group C+D.

4.7.3 Comparison of original and recalibrated cut-off parameters

To further characterize the shift in classification observed after recalibration, the coefficients of the original published HD-ISS Stage 1 cut-off curves were compared with those estimated on the harmonized reference sample. As shown in Table 4.13, recalibration modified both intercept and slope for the caudate and putamen curves, consistent with the redistribution of

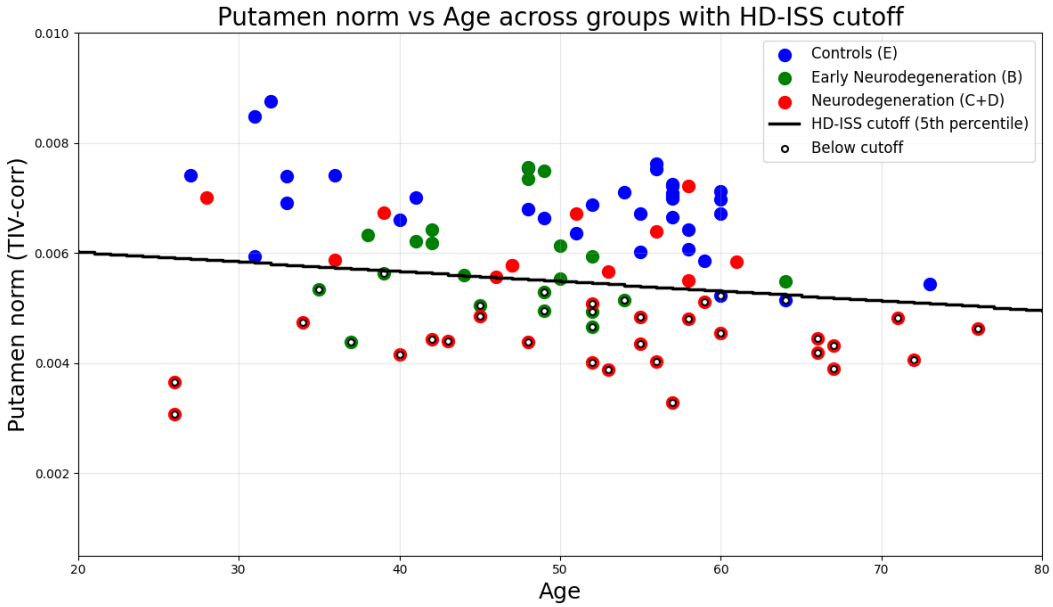


(a) Original published thresholds

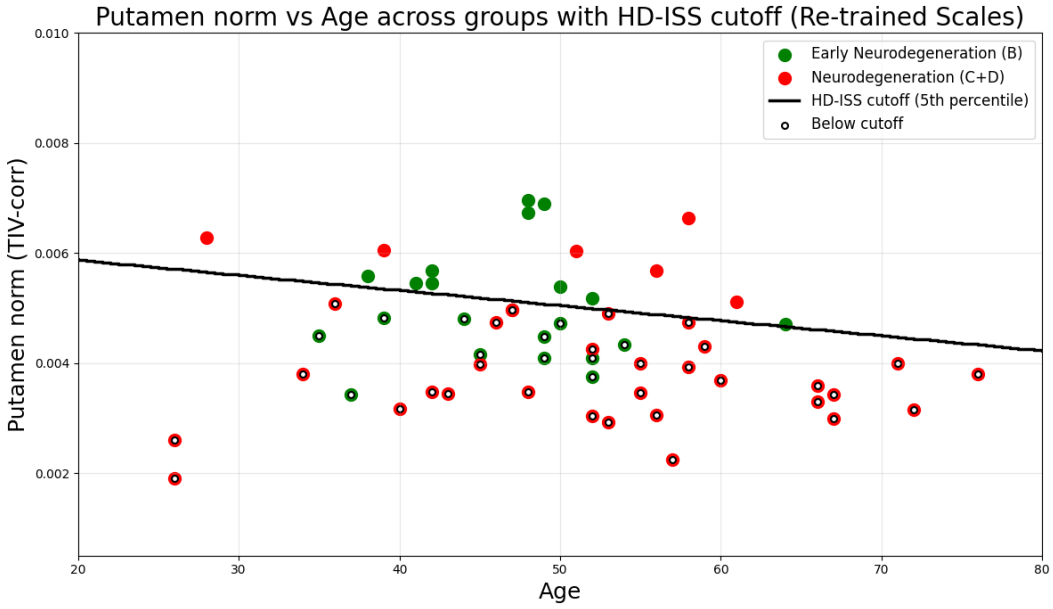


(b) Recalibrated pipeline

Figure 4.11. Comparison between the direct application of the original published HD-ISS Stage 1 cut-offs and the recalibrated pipeline for normalized caudate volume in the *iMarkHD* sample. In both panels, normalized caudate volume is plotted against age.



(a) Original published thresholds



(b) Recalibrated pipeline

Figure 4.12. Comparison between the direct application of the original published HD-ISS Stage 1 cut-offs and the recalibrated pipeline for normalized putamen volume in the *iMarkHD* sample. In both panels, normalized putamen volume is plotted against age.

subject-level observations relative to the classification boundaries shown in Figures 4.11 and 4.12.

Table 4.13. Comparison between the original published age-dependent HD-ISS Stage 1 cut-off parameters and the recalibrated parameters estimated on the harmonized reference sample. Cut-offs were modeled as cutoff(age) = $a + b \cdot age$, where a is the intercept and b is the slope.

	Original published HD-ISS		Recalibrated on harmonized reference	
ROI	Intercept (a)	Slope (b)	Intercept (a)	Slope (b)
Caudate	0.004558000	-1.3000×10^{-5}	0.004177644	-9.620×10^{-6}
Putamen	0.006364000	-1.8000×10^{-5}	0.006419058	-2.7492×10^{-5}

Overall, these findings indicate that the direct application of the published HD-ISS Stage 1 thresholds to the independent *iMarkHD* sample captured a monotonic increase in striatal abnormality across clinical groups, whereas the recalibrated pipeline yielded a higher proportion of subjects classified below the age-adjusted cut-offs, particularly in the group with more advanced neurodegeneration. This change was accompanied by a measurable repositioning of the age-dependent classification curves, especially for the putamen.

5 Discussion

5.1 Main interpretation of the findings

This thesis addressed the translational question, namely whether quantitative neuroimaging-based cutoff scales remain stable and interpretable when transferred to independent multi-site settings. Using the HD-ISS Stage 1 framework as a case study, the results show that this portability cannot be assumed a priori. Direct application of the published thresholds to the external normative cohort produced inflated abnormality rates and substantial heterogeneity across datasets (Figure 4.1; Table 4.1), indicating that the practical usability of imaging-based staging rules depends not only on the biological relevance of the underlying biomarker, but also on the technical conditions under which it is acquired, processed, and interpreted [4].

A second general conclusion is that site-related variability represented a major obstacle to threshold portability. As shown by the raw structural MRI analyses (Table 4.2; Figure 4.3), the problem was not limited to caudate and putamen, but reflected a broader instability of quantitative measurements across datasets. In this setting, harmonization was necessary but not sufficient: ComBat was the most effective method for reducing site effects (Table 4.3), yet applying the original HD-ISS cut-offs after harmonization alone did not fully restore the intended normative behaviour (Figure 4.6; Table 4.4). The best overall solution was therefore obtained by combining har-

monization and external recalibration, with the harmonized-to-harmonized configuration showing the most stable out-of-sample behaviour (Figure 4.7; Table 4.5).

5.1.1 Portability of HD-ISS Stage 1 cut-offs in external multi-site data

The direct transfer of the original HD-ISS Stage 1 thresholds to the external multi-site cohort was the clearest demonstration that biologically meaningful MRI cut-offs do not automatically remain portable outside the context in which they were originally derived. Because the Stage 1 framework relies on age-dependent volumetric cut-offs for caudate and putamen combined through an either-or rule [4], even relatively modest technical shifts at the level of the individual landmarks may translate into substantial changes in final classification. This helps explain the widespread inflation of abnormal classifications observed in Section 4.1.

The site-effect analyses further clarify that this was not a problem specific to one single scale, but the expression of a more general methodological instability. Once a continuous quantitative biomarker is converted into a threshold-based rule, any systematic measurement shift may produce a disproportionate effect on classification. In this sense, the HD-ISS Stage 1 framework represents a particularly informative example of a broader translational challenge affecting quantitative neuroimaging-based scales.

The harmonization results showed that an important part of this miscalibration was attributable to technical variability between datasets. However, the persistence of elevated classification rates even after ComBat harmonization indicates that the problem did not concern the feature space alone. The published cut-offs were not simply being applied to technically distorted data, but were also being transferred from a reference context that was not fully matched to the present cohort. This explains why harmonization alone was insufficient and why a subsequent recalibration step became necessary.

5.1.2 External recalibration as the key methodological contribution

The recalibration analyses identify the main methodological contribution of the present work. As shown in Figure 4.7 and Table 4.5, external recalibration improved threshold behaviour in all evaluated scenarios, but the most coherent results were obtained when both threshold estimation and threshold application were performed in the same harmonized space. In particular, the harmonized-to-harmonized configuration showed the smallest discrepancy between training and test classifications and the narrowest variability across held-out datasets.

This finding suggests that harmonization and recalibration act on complementary aspects of the problem. Harmonization reduces technical inconsistency in the measurements, whereas recalibration adapts the age-dependent thresholds to the external normative context in which the framework is being redeployed. The recalibrated parameters reported in Tables 4.6 and 4.7 support this interpretation, showing that external adaptation affected not only the intercepts but also the age dependence of the caudate and putamen cut-off curves.

5.1.3 Translational relevance: application to the *iMarkHD* sample

The application to the *iMarkHD* sample adds a first translational validation layer to the study. As shown in Table 4.11, the direct application of the original HD-ISS thresholds already captured the expected monotonic increase in striatal abnormality across the three clinical groups, indicating that the original framework retained biological sensitivity in the patient sample. However, the recalibrated pipeline modified the classification profile in a direction that was more consistent with the expected disease-severity gradient.

In particular, the recalibrated cut-offs increased the proportion of subjects below threshold in both patient groups, with the clearest effect in the neurodegeneration group (C+D), especially for putamen-related abnormality and for the final N+ classification (Table 4.12). The subject-level plots (Figures 4.11 and 4.12) and the parameter changes reported in Table 4.13 fur-

ther indicate that recalibration repositioned the classification boundaries in a way that was coherent with the observed patient distributions. Although this application remains preliminary, it suggests that the methodological improvements identified in the normative analyses may also be relevant for disease-related classification in patient cohorts.

5.1.4 Normative modelling without explicit harmonization

Normative modelling was included as a conceptually important alternative to explicit harmonization. In principle, a hierarchical normative model could account for scanner- and site-related variability directly during estimation and thus provide a flexible framework for external deployment. In the present study, however, this approach only partially addressed the portability problem. As shown in Tables 4.8 and 4.9, and in Figure 4.8, external-site calibration remained suboptimal across the healthy test cohorts, with the clearest deviations observed for the putamen and for the combined OR criterion.

The centile plots (Figures 4.9 and 4.10) suggest that the HBR models were able to capture the expected age-related decline in striatal volumes, but not to eliminate all residual dataset-dependent displacement. Therefore, while normative modelling proved informative and partially corrective, it did not match the robustness of the best harmonization-plus-recalibration solution. In this specific multi-site application, it should thus be considered a complementary strategy rather than a replacement for explicit harmonization when the goal is to maximize the external portability of HD-ISS Stage 1 cut-offs.

5.1.5 Limitations of the study

FreeSurfer version mismatch

An important limitation of the present study concerns the mismatch between the software version used to generate the external normative dataset and the version underlying the original HD-ISS framework. The 5k+ normative cohort analysed in this thesis was processed with FreeSurfer 7.3.2, whereas the published HD-ISS Stage 1 cut-offs were derived from volumes obtained with FreeSurfer 6.0.1 [4]. This is not a negligible technical detail,

because the -ISS Stage 1 framework is based specifically on intracranial-volume-adjusted caudate and putamen volumes.

This concern is supported by the study of Knights et al., who directly compared FreeSurfer 6.0.1 and FreeSurfer 7.2.0 in the HD-YAS cohort and showed that the software update produces substantial differences in adjusted striatal volumes, with FreeSurfer 7 yielding systematically higher adjusted values than FreeSurfer 6 [37]. The magnitude of the reported difference is summarised in Table 5.1. Because these measures correspond exactly to the volumetric quantities used by HD-ISS Stage 1, such inter-version bias has the potential to alter the behaviour of the published cut-offs when they are transferred to external data.

Table 5.1. Reported differences in intracranial-volume-adjusted striatal volumes between FreeSurfer 6.0.1 and 7.2.0 from Knights et al. Values are expressed as percentage increase in FS7 relative to FS6 [37].

Measure	FS7 vs FS6 variation
Adjusted caudate	+7.0%
Adjusted putamen	+8.2%

At the same time, this limitation should be interpreted with caution. The comparison reported by Knights et al. was performed between FreeSurfer 6.0.1 and 7.2.0, not between 6.0.1 and 7.3.2, and therefore does not provide a correction factor fully matched to the processing pipeline used in the present study [37]. More generally, previous work has shown that compatibility across FreeSurfer versions is not uniform and that inter-version differences may affect downstream analyses [38]. In practical terms, if the correction proposed by Knights et al. had been applied here, the results would have become even worse already at the initial normative step, with a further increase in the percentage of subjects falling below threshold. For this reason, that correction was not included in the final pipeline.

Nevertheless, the mismatch between FreeSurfer versions remains an important limitation, especially in the first analysis, where the original HD-ISS thresholds were directly transferred to the external normative cohort. In this setting, part of the observed miscalibration may reflect not only site-related variability, but also an upstream incompatibility between the segmentation framework used to derive the published Stage 1 cut-offs and the one used to generate the normative data analysed in this thesis. However,

the fact that performance improved after recalibration, and even more after harmonization-based recalibration, indicates that these processing-related differences can be at least partially compensated for, although they are unlikely to be completely eliminated.

Differences between training and test pools

A second limitation concerns the difference between the training and test pools used for external recalibration and normative modelling. The leave-sites-out design adopted in this thesis was methodologically appropriate, because it allowed the transportability of the proposed pipeline to be evaluated under realistic out-of-sample conditions rather than within a single pooled cohort. However, this design also implies that the training and test sets were not fully interchangeable in terms of demographic composition, acquisition context, and statistical structure.

This mismatch was reflected in the results, most clearly in the behaviour of HCP-YoungAdult, which repeatedly showed the largest deviation from the expected normative profile across multiple analyses, including the direct application of the original thresholds (Figure 4.1), the harmonized reapplication of the published cut-offs (Figure 4.6), the recalibration scenarios (Figure 4.7), and the normative modelling results (Table 4.8; Figure 4.8). This pattern suggests that part of the residual instability observed out of sample may depend not only on the thresholding framework itself, but also on the limited representativeness of the training pool relative to some of the held-out target datasets.

Therefore, although the train-test separation was essential for evaluating external portability, it also constitutes a limitation in the sense that the normative reference used for threshold estimation did not fully span the variability encountered in the external cohorts. In practical terms, this means that the performance of the recalibrated pipeline should be interpreted as conditional on the composition of the reference sample. A broader and more heterogeneous training pool might further improve transportability, especially toward external sites with acquisition or demographic properties that are only partially represented in the current dataset.

Limited patient-level external validation

A final limitation concerns the still limited external validation at the patient level. The ultimate goal of this line of research is not simply to improve normative calibration in healthy control cohorts, but to obtain a pipeline that classifies patient data more robustly and meaningfully across sites. In the present study, this aspect was addressed only partially through the application to the *iMarkHD* sample, which provided an initial indication that the recalibrated pipeline may improve the identification of striatal abnormality in patient groups (Tables 4.11 and 4.12; Figures 4.11 and 4.12).

However, this analysis does not yet constitute a full external validation of patient-level classification. The *iMarkHD* sample remains relatively limited in size, and its use in the present work should be interpreted primarily as a proof of concept rather than as definitive evidence of generalizability. Moreover, the observed gain was assessed in a specific cohort and within a specific acquisition and processing context, so it remains unclear to what extent the same improvement would be preserved across additional HD datasets, new sites, or broader clinical spectra.

For this reason, the present findings should be viewed as an encouraging but still preliminary step toward patient-level translation. Further studies on independent HD cohorts, ideally acquired across multiple external sites, will be needed to determine whether the harmonization-plus-recalibration pipeline consistently improves clinically meaningful classification beyond the current sample.

Taken together, these limitations indicate that the present findings should be interpreted as a strong methodological proof of concept rather than as a fully definitive solution to the portability problem. Even so, the study shows that the external applicability of quantitative neuroimaging-based cutoff scales can be improved in a meaningful way when technical variability is addressed explicitly, and it provides a concrete framework on which broader future validation efforts can build.

6

Conclusions

This thesis addressed a central translational issue in quantitative neuroimaging: whether imaging-derived cutoff scales remain stable and interpretable when transferred across independent multi-site datasets. Using the HD-ISS Stage 1 framework as a case study, the results showed that this portability cannot be assumed a priori. The direct application of the published MRI-based thresholds to an external normative cohort produced inflated abnormality rates and marked variability across datasets, indicating that biologically meaningful imaging biomarkers do not automatically translate into robust and portable classification rules.

The analyses further showed that site-related variability represented a major obstacle to the external applicability of the original cut-offs. Among the evaluated strategies, the most effective solution was the combination of harmonization and external recalibration. ComBat provided the best reduction of site effects, but harmonization alone was not sufficient to restore the expected behaviour of the original HD-ISS thresholds. The best overall performance was obtained when threshold estimation and application were both performed in the same harmonized space. By contrast, normative modelling without explicit harmonization provided only a partial correction and did not reach the same level of robustness.

The application to the *iMarkHD* sample added an initial translational perspective to the study, suggesting that the recalibrated pipeline may also improve the identification of striatal abnormality in patient data, particularly in the more advanced neurodegeneration group. Although still pre-

liminary, this result supports the relevance of the proposed methodological framework beyond normative calibration alone.

Overall, this work shows that the external usability of quantitative neuroimaging-based cutoff scales depends critically on how technical variability is handled across acquisition, processing, and threshold definition. In this study, harmonization followed by external recalibration provided the most robust strategy for improving the portability of HD-ISS Stage 1 cut-offs in multi-site data.

6.1 Future perspectives

Future work should validate the proposed pipeline in additional independent HD cohorts, ideally acquired across multiple external sites. Further investigation is also needed to clarify the impact of software-version differences on volumetric thresholds and to determine whether more refined normative modelling approaches can improve external calibration. More broadly, extending similar analyses to other MRI-based staging systems may help define general methodological principles for the reliable translation of quantitative neuroimaging into multi-site applications.

References

[1] G. Saltarelli et al. “Quantitative MRI in Neuroimaging: A Review of Techniques, Biomarkers, and Emerging Clinical Applications”. In: *Brain Sciences* 15.10 (2025), p. 1088. DOI: 10.3390/brainsci15101088.

[2] Jakub Marek et al. “Quantitative brain volumetry in neurological disorders: from disease mechanisms to software solutions”. In: *Polish Journal of Radiology* 90 (2025), e299–e306. DOI: 10.5114/pjr/203781.

[3] A. Giorgio and N. De Stefano. “Clinical use of brain volumetry”. In: *Journal of Magnetic Resonance Imaging* 37.1 (2013), pp. 1–14. DOI: 10.1002/jmri.23671.

[4] S. J. Tabrizi et al. “A biological classification of Huntington’s disease: the Integrated Staging System”. In: *The Lancet Neurology* 21.7 (2022). On behalf of the Huntington’s Disease Regulatory Science Consortium (HD-RSC), pp. 632–644. DOI: 10.1016/S1474-4422(22)00120-X.

[5] C. A. Ross et al. “Huntington disease: natural history, biomarkers and prospects for therapeutics”. In: *Nature Reviews Neurology* 10.4 (2014), pp. 204–216. DOI: 10.1038/nrneurol.2014.24.

[6] T. van Eimeren et al. “Neuroimaging biomarkers in Huntington’s disease: Preparing for a new era of therapeutic development”. In: *Parkinsonism & Related Disorders* 114 (2023), p. 105488. DOI: 10.1016/j.parkreldis.2023.105488.

[7] J. S. Paulsen et al. “Detection of Huntington’s disease decades before diagnosis: the Predict-HD study”. In: *Journal of Neurology, Neurosurgery & Psychiatry* 79.8 (2008), pp. 874–880. DOI: 10.1136/jnnp.2007.128728.

[8] K. M. Kinnunen et al. “Volumetric MRI-Based Biomarkers in Huntington’s Disease: An Evidentiary Review”. In: *Frontiers in Neurology* 12 (2021). On behalf of the Huntington’s Disease Regulatory Science Consortium (HD-RSC), p. 712555. DOI: 10.3389/fneur.2021.712555.

[9] R. I. Scahill et al. “Biological and clinical characteristics of gene carriers far from predicted onset in the Huntington’s disease Young Adult Study (HD-YAS): a cross-sectional analysis”. In: *The Lancet Neurology* 19.6 (2020), pp. 502–512. DOI: 10.1016/S1474-4422(20)30143-5.

[10] J. P. Fortin et al. “Harmonization of cortical thickness measurements across scanners and sites”. In: *NeuroImage* 167 (2018), pp. 104–120. DOI: 10.1016/j.neuroimage.2017.11.024.

[11] J. M. M. Bayer et al. “Site effects how-to and when: An overview of retrospective techniques to accommodate site effects in multi-site neuroimaging analyses”. In: *Frontiers in Neurology* 13 (2022), p. 923988. DOI: 10.3389/fneur.2022.923988.

[12] C. Marzi et al. “Efficacy of MRI data harmonization in the age of machine learning: a multicenter study across 36 datasets”. In: *Scientific Data* 11.1 (2024), p. 115. DOI: 10.1038/s41597-023-02421-7.

[13] Andrew A. Chen et al. “Mitigating site effects in covariance for machine learning in neuroimaging data”. In: *Human Brain Mapping* 43.4 (2022). for the Alzheimer’s Disease Neuroimaging Initiative, pp. 1179–1195. DOI: 10.1002/hbm.25688.

[14] R. Pomponio et al. “Harmonization of large MRI datasets for the analysis of brain imaging patterns throughout the lifespan”. In: *NeuroImage* 208 (2020), p. 116450. DOI: 10.1016/j.neuroimage.2019.116450.

[15] J. P. Fortin et al. “Harmonization of multi-site diffusion tensor imaging data”. In: *NeuroImage* 161 (2017), pp. 149–170. DOI: 10.1016/j.neuroimage.2017.08.047.

[16] A. F. Marquand et al. “Conceptualizing mental disorders as deviations from normative functioning”. In: *Molecular Psychiatry* 24.10 (2019). Erratum published in Molecular Psychiatry, 2019, 24(10):1565. doi:10.1038/s41380-019-0451-z, pp. 1415–1424. DOI: 10.1038/s41380-019-0441-1.

[17] S. Rutherford et al. “The normative modeling framework for computational psychiatry”. In: *Nature Protocols* 17.7 (2022), pp. 1711–1734. DOI: 10.1038/s41596-022-00696-5.

[18] S. M. Kia et al. “Hierarchical Bayesian Regression for Multi-Site Normative Modeling of Neuroimaging Data”. In: *arXiv* (2020). Preprint. arXiv: 2005.12055 [stat.ML].

[19] C. J. Fraza et al. “The Extremes of Normative Modelling”. In: *bioRxiv* (2022). Preprint. DOI: 10.1101/2022.08.23.505049.

[20] A. A. A. de Boer et al. “Non-Gaussian normative modelling with hierarchical Bayesian regression”. In: *Imaging Neuroscience* 2 (2024). DOI: 10.1162/imag_a_00132.

[21] O. Esteban et al. “MRIQC: Advancing the automatic prediction of image quality in MRI from unseen sites”. In: *PLoS ONE* 12.9 (2017), e0184661. DOI: 10.1371/journal.pone.0184661.

[22] Oscar Esteban et al. “fMRIPrep: a robust preprocessing pipeline for functional MRI”. In: *Nature Methods* 16.1 (2019), pp. 111–116. DOI: 10.1038/s41592-018-0235-4.

[23] N. Mattsson-Carlgren et al. “The implications of different approaches to define AT(N) in Alzheimer disease”. In: *Neurology* 94.21 (2020), e2233–e2244. DOI: 10.1212/WNL.0000000000009485.

[24] D. C. Van Essen et al. “The WU-Minn Human Connectome Project: an overview”. In: *NeuroImage* 80 (2013). for the WU-Minn HCP Consortium, pp. 62–79. DOI: 10.1016/j.neuroimage.2013.05.041.

[25] M. F. Glasser et al. “The minimal preprocessing pipelines for the Human Connectome Project”. In: *NeuroImage* 80 (2013), pp. 105–124. DOI: 10.1016/j.neuroimage.2013.04.127.

[26] Russell A. Poldrack et al. “A phenome-wide examination of neural and cognitive function”. In: *Scientific Data* 3 (2016), p. 160110. DOI: 10.1038/sdata.2016.110.

[27] Dongtao Wei et al. “Structural and functional brain scans from the cross-sectional Southwest University adult lifespan dataset”. In: *Scientific Data* 5 (2018), p. 180134. DOI: 10.1038/sdata.2018.134.

[28] Kate B. Nooner et al. "The NKI-Rockland Sample: A model for accelerating the pace of discovery science in psychiatry". In: *Frontiers in Neuroscience* 6 (2012), p. 152. DOI: 10.3389/fnins.2012.00152.

[29] Jason R. Taylor et al. "The Cambridge Centre for Ageing and Neuroscience (Cam-CAN) data repository: Structural and functional MRI, MEG, and cognitive data from a cross-sectional adult lifespan sample". In: *NeuroImage* 144.Pt B (2017), pp. 262–269. DOI: 10.1016/j.neuroimage.2015.09.018.

[30] Meredith A. Shafto et al. "The Cambridge Centre for Ageing and Neuroscience (Cam-CAN) study protocol: a cross-sectional, lifespan, multidisciplinary examination of healthy cognitive ageing". In: *BMC Neurology* 14 (2014), p. 204. DOI: 10.1186/s12883-014-0204-1.

[31] Susan Y. Bookheimer et al. "The Lifespan Human Connectome Project in Aging: An overview". In: *NeuroImage* 185 (2019), pp. 335–348. DOI: 10.1016/j.neuroimage.2018.10.009.

[32] Michael P. Harms et al. "Extending the Human Connectome Project across ages: Imaging protocols for the Lifespan Development and Aging projects". In: *NeuroImage* 183 (2018), pp. 972–984. DOI: 10.1016/j.neuroimage.2018.09.060.

[33] Denise C. Park et al. "The Dallas Lifespan Brain Study: A Comprehensive Adult Lifespan Data Set of Brain and Cognitive Aging". In: *Scientific Data* 12 (2025), p. 846. DOI: 10.1038/s41597-025-04847-7.

[34] Jennifer Tremblay-Mercier et al. "Open science datasets from PREVENT-AD, a longitudinal cohort of pre-symptomatic Alzheimer's disease". In: *NeuroImage: Clinical* 31 (2021), p. 102733. DOI: 10.1016/j.nicl.2021.102733.

[35] Pamela J. LaMontagne et al. "OASIS-3: Longitudinal Neuroimaging, Clinical, and Cognitive Dataset for Normal Aging and Alzheimer Disease". In: *medRxiv* (2019). DOI: 10.1101/2019.12.13.19014902.

[36] Daniel J. van Wamelen et al. "Study protocol for the iMarkHD study in individuals with Huntington's disease". In: *Journal of Huntington's Disease* (2024), pp. 1–11. DOI: 10.1177/18796397241288165.

[37] Harry Knights et al. “Freesurfer Software Update Significantly Impacts Striatal Volumes in the Huntington’s Disease Young Adult Study and Will Influence HD-ISS Staging”. In: *Journal of Huntington’s Disease* 13 (2024). and the HD-YAS investigators, pp. 77–90. DOI: 10.3233/JHD-231512.

[38] Elizabeth Haddad et al. “Multisite test–retest reliability and compatibility of brain metrics derived from FreeSurfer versions 7.1, 6.0, and 5.3”. In: *Human Brain Mapping* 44.4 (2023), pp. 1515–1532. DOI: 10.1002/hbm.26147.

Appendix

Feature Distributions and Post-harmonization ANCOVA Tables

This section reports the dataset-wise feature distributions before harmonization and after each evaluated harmonization strategy, together with the corresponding ANCOVA results used to quantify residual site effects.

Pre-harmonization feature distributions

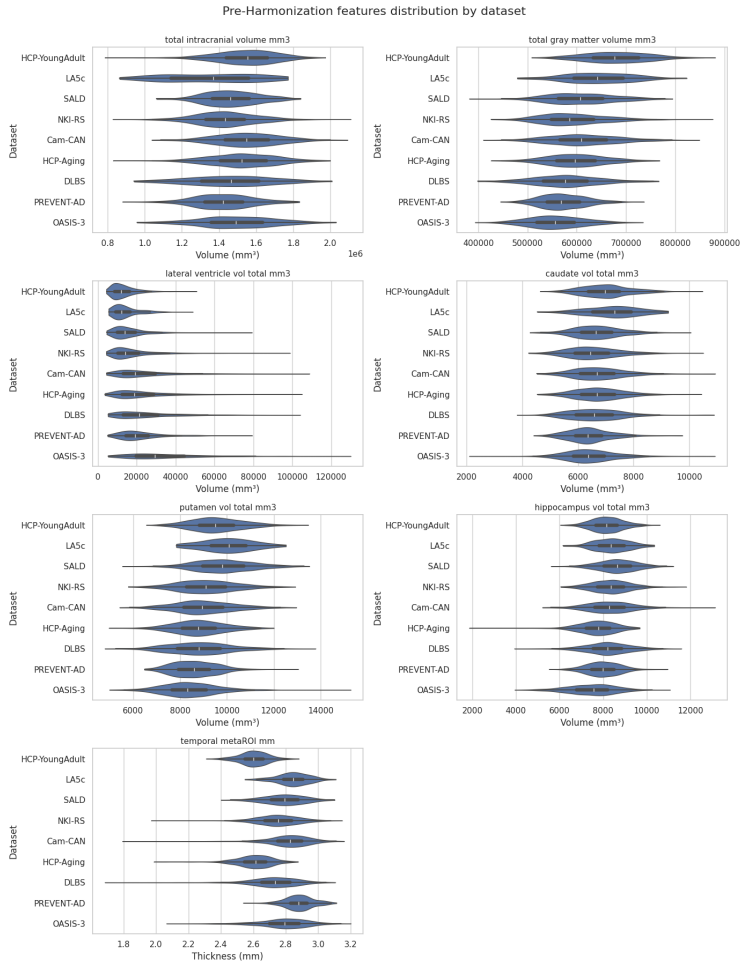


Figure 6.1. Feature distributions across datasets before harmonization. Violin plots are shown for all structural MRI features included in the harmonization analysis.

ComBat

Table 6.1. ANCOVA results for the structural MRI features after ComBat harmonization. For each feature, the table reports the test statistic and p-value associated with the site effect, together with its effect size (η^2), and the regression coefficients and p-values for age and sex.

ROI	F	p_{site}	η^2_{site}	β_{age}	p_{age}	β_{sex}	p_{sex}
Total intracranial volume	0.0575	0.999903	0.000063	85.2566	0.5645401	198384.16	0.00E+00
Total gray matter volume	0.0126	1.000000	0.000011	-2145.97	0.00E+00	59908.661	0.00E+00
Lateral ventricle volume	0.6121	0.768458	0.000732	416.4211	1.13E-257	5510.0581	2.61E-62
Caudate volume	0.0074	1.000000	0.000010	-13.0938	8.51E-52	541.3704	1.83E-104
Putamen volume	0.0068	1.000000	0.000007	-39.7521	0.00E+00	850.8896	1.05E-193
Hippocampus volume	0.0404	0.999975	0.000048	-26.6949	2.93E-222	505.4245	1.10E-104
Temporal meta-ROI thickness	0.0361	0.999984	0.000047	-0.00364	3.25E-201	0.000871	0.7903989

Note. F and p_{site} refer to the dataset term, while η^2_{site} quantifies the magnitude of the site effect. β_{age} and β_{sex} denote the regression coefficients for age and sex, respectively, with their associated p-values.

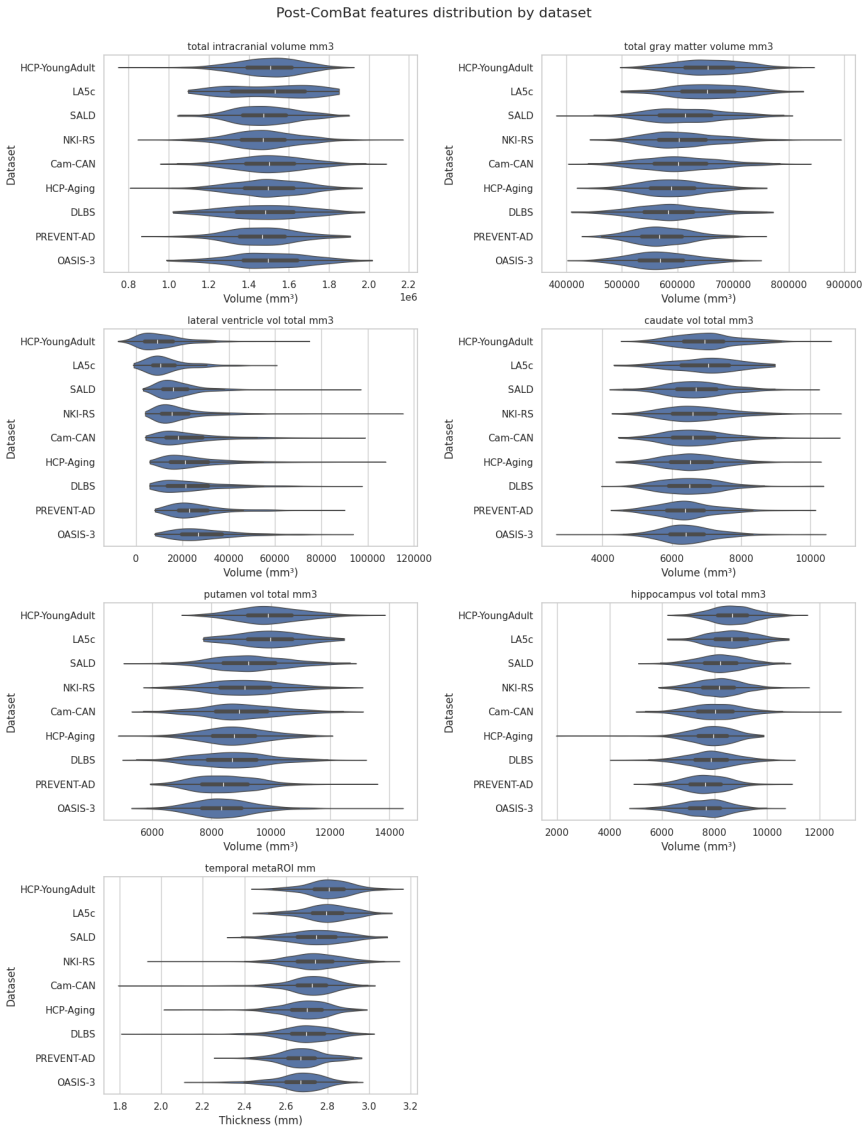


Figure 6.2. Feature distributions across datasets after ComBat harmonization.

ComBat-GAM

Table 6.2. ANCOVA results for the structural MRI features after ComBat-GAM harmonization.

ROI	F	p_{site}	η^2_{site}	β_{age}	p_{age}	β_{sex}	p_{sex}
Total intracranial volume	1.3391	2.19E-01	0.001467	101.8772	4.91E-01	198443.88	0.00E+00
Total gray matter volume	0.2337	9.85E-01	0.000202	-2146.0650	0.00E+00	59919.0941	0.00E+00
Lateral ventricle volume	8.1967	4.42E-11	0.009458	443.3264	1.70E-288	5544.0275	3.14E-63
Caudate volume	2.2166	2.35E-02	0.003051	-12.9015	1.98E-50	541.3360	1.41E-104
Putamen volume	0.1146	9.99E-01	0.000118	-39.7530	0.00E+00	850.9175	7.84E-194
Hippocampus volume	4.7761	7.25E-06	0.005657	-27.0165	2.99E-227	504.7294	1.71E-104
Temporal meta-ROI thickness	2.4889	1.08E-02	0.003254	-0.00368	3.86E-205	0.000588	8.57E-01

Note. F and p_{site} refer to the dataset term, while η^2_{site} quantifies the magnitude of the site effect. β_{age} and β_{sex} denote the regression coefficients for age and sex, respectively, with their associated p-values.

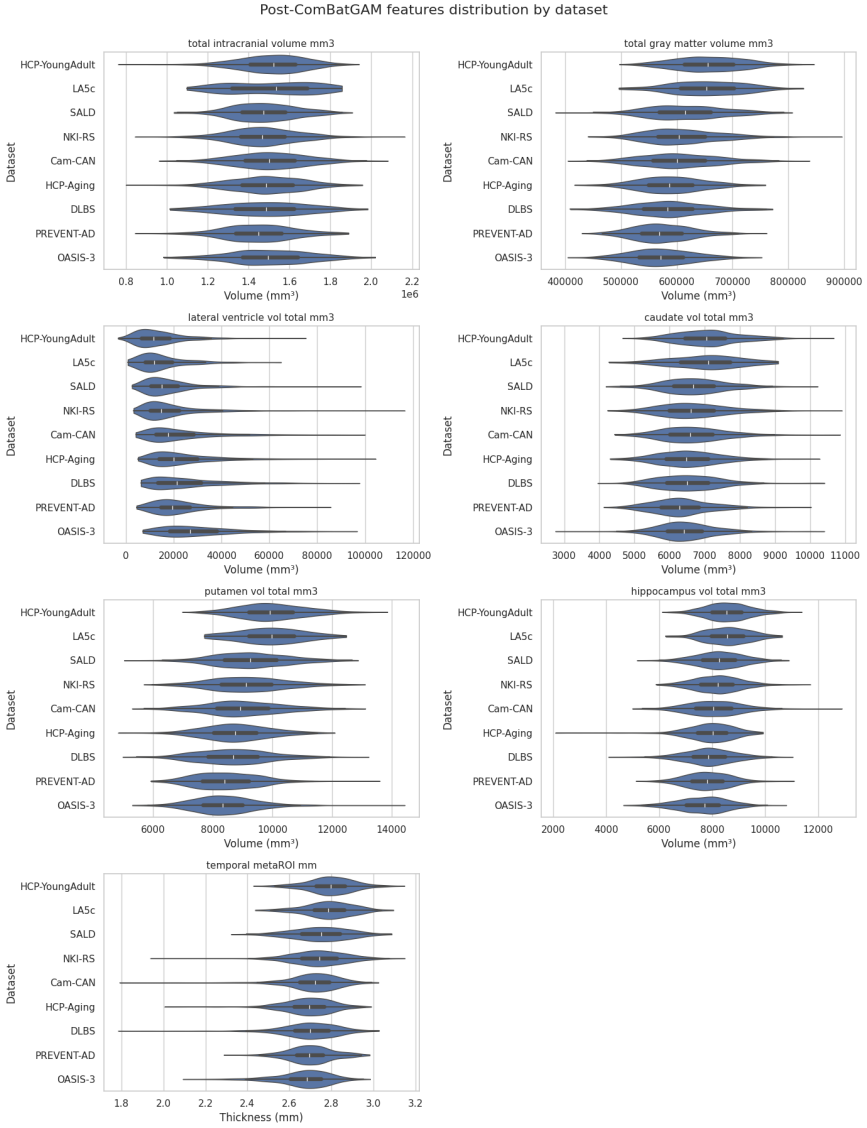


Figure 6.3. Feature distributions across datasets after ComBat-GAM harmonization.

CovBat

Table 6.3. ANCOVA results for the structural MRI features after CovBat harmonization.

ROI	F	p_{site}	η^2_{site}	β_{age}	p_{age}	β_{sex}	p_{sex}
Total intracranial volume	0.0623	0.999868	0.000069	40.8332	0.7818703	197284.33	0.00E+00
Total gray matter volume	0.0089	1.000000	0.000008	-2138.08	0.00E+00	60071.685	0.00E+00
Lateral ventricle volume	0.9216	0.497087	0.001116	407.6029	8.29E-247	5314.085	7.11E-58
Caudate volume	0.0139	1.000000	0.000019	-13.3070	1.79E-53	536.5495	7.22E-103
Putamen volume	0.0029	1.000000	0.000003	-39.5768	0.00E+00	856.1064	4.03E-196
Hippocampus volume	0.0632	0.999860	0.000076	-26.3053	1.10E-217	515.3734	3.24E-109
Temporal meta-ROI thickness	0.1025	0.999149	0.000136	-0.00356	8.07E-194	0.002674	0.4135148

Note. F and p_{site} refer to the dataset term, while η^2_{site} quantifies the magnitude of the site effect. β_{age} and β_{sex} denote the regression coefficients for age and sex, respectively, with their associated p-values.

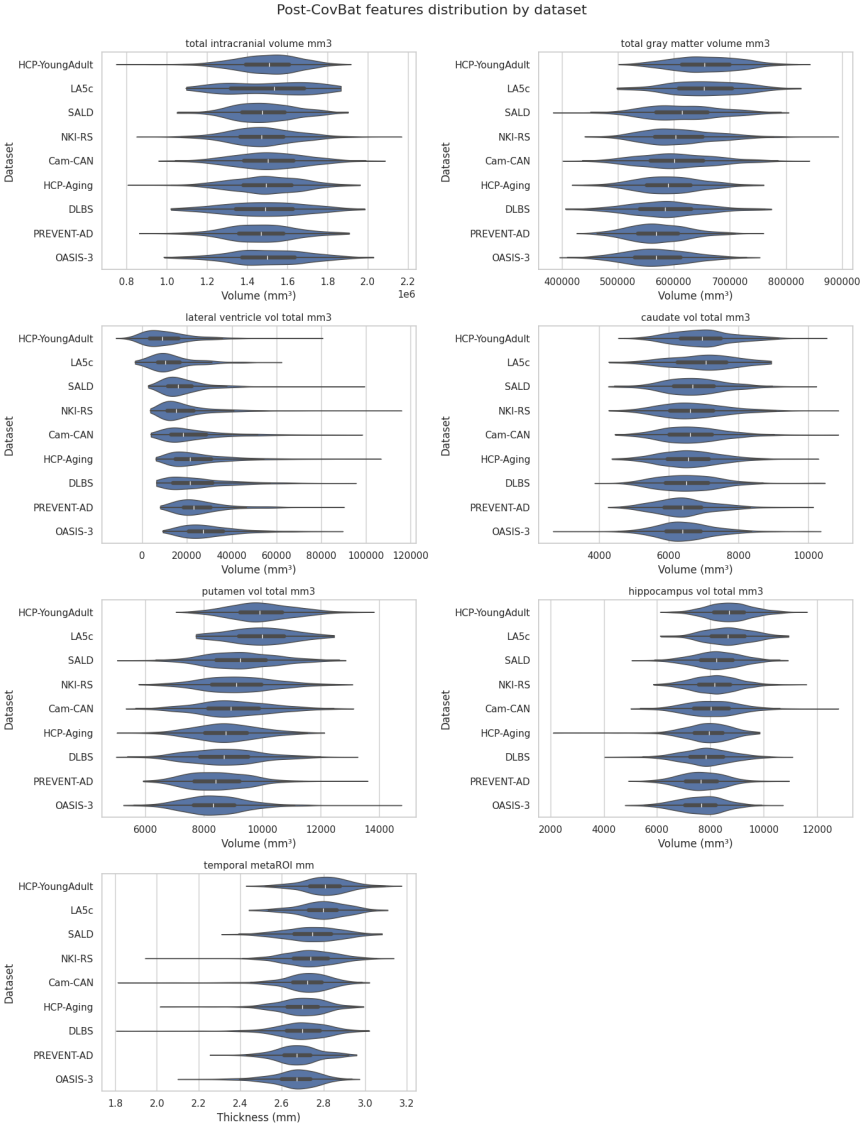
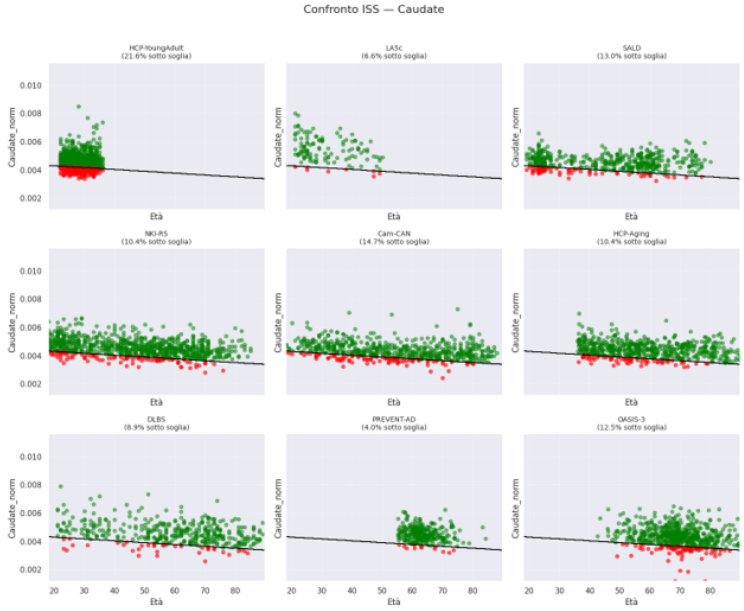


Figure 6.4. Feature distributions across datasets after CovBat harmonization.

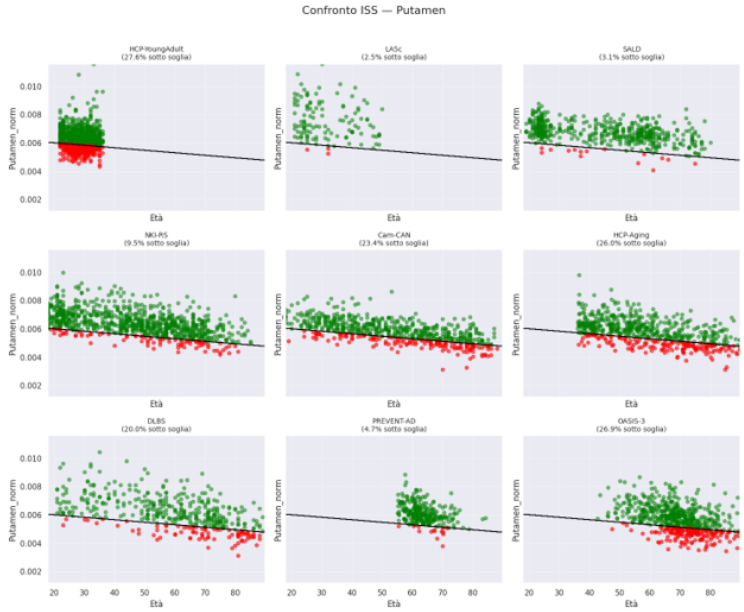
Application of the Published HD-ISS Stage 1 Thresholds to Normative Data

Across the following figures, each panel corresponds to one dataset and displays TIV-normalized regional volume as a function of age. The black line represents the published age-dependent threshold, subjects below the threshold are shown in red, and subjects above it are shown in green. The percentage reported in each panel indicates the proportion of normative subjects classified below the corresponding threshold.

Unharmonized normative data



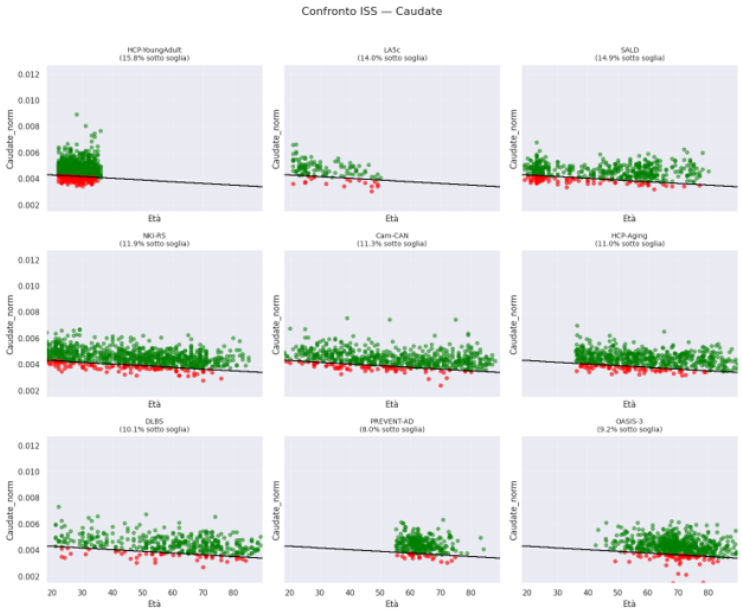
(a) Application of the published HD-ISS Stage 1 caudate threshold to the unharmonized normative datasets.



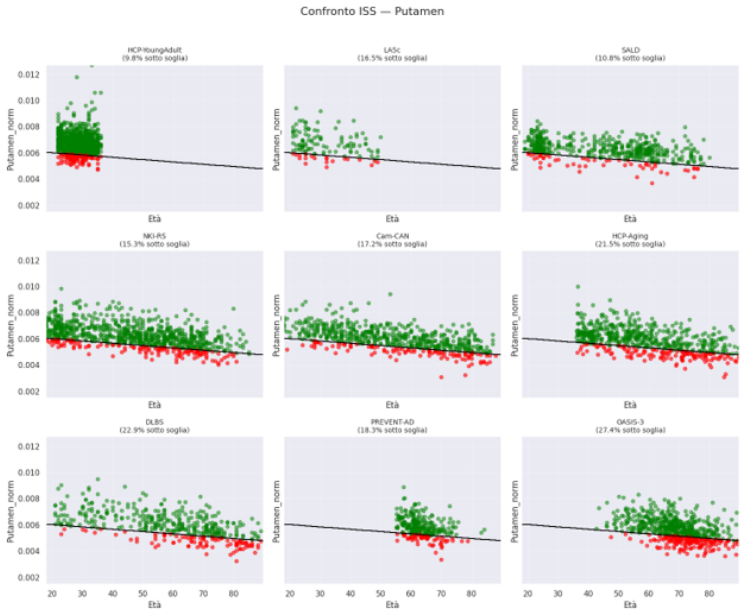
(b) Application of the published HD-ISS Stage 1 putamen threshold to the unharmonized normative datasets.

Figure 6.5. Application of the published HD-ISS Stage 1 age-dependent thresholds to the original unharmonized multi-site normative datasets, shown separately for the caudate (top) and putamen (bottom).

Harmonized normative data



(a) Application of the published HD-ISS Stage 1 caudate threshold to the harmonized normative datasets.



(b) Application of the published HD-ISS Stage 1 putamen threshold to the harmonized normative datasets.

Figure 6.6. Application of the published HD-ISS Stage 1 age-dependent thresholds to the harmonized multi-site normative datasets, shown separately for the caudate (top) and putamen (bottom).

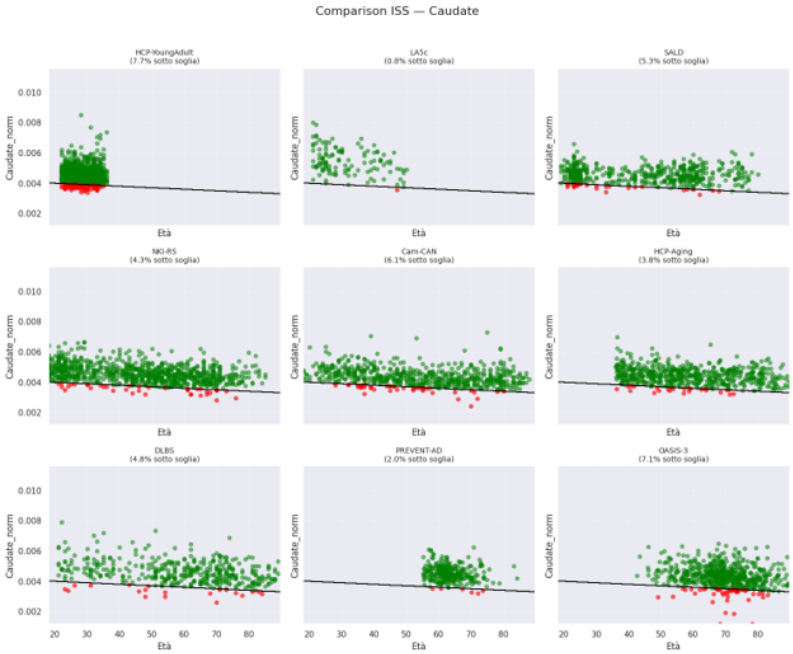
External Recalibration of the HD-ISS Stage 1 Thresholds

Across the following figures, thresholds were estimated on the training set corresponding to each configuration and then applied to the respective independent test datasets. Each panel corresponds to one test dataset and displays TIV-normalized regional volume as a function of age. The black line represents the retrained age-dependent threshold, subjects below the threshold are shown in red, and subjects above it are shown in green. The percentage reported in each panel indicates the proportion of subjects classified below the corresponding threshold.

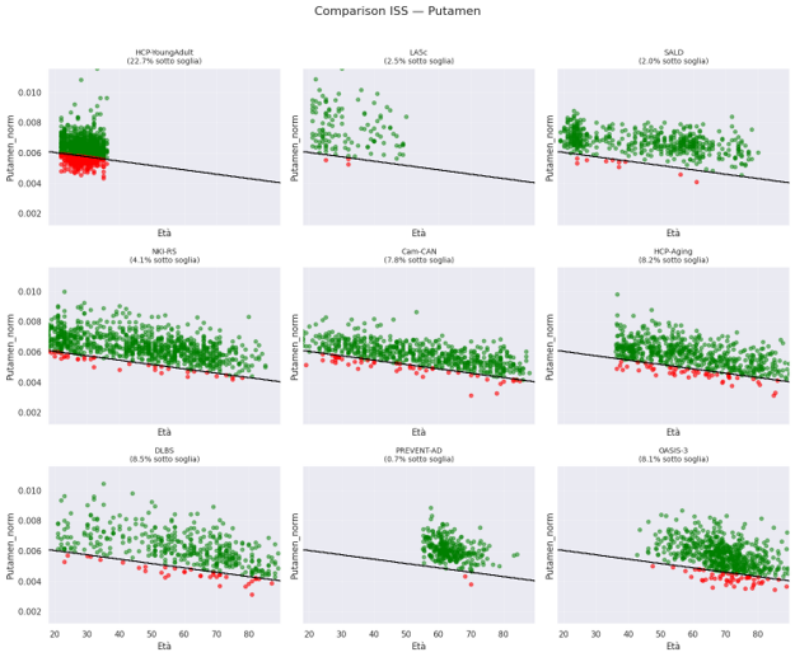
Raw-to-raw configuration

Table 6.4. Application of the retrained HD-ISS Stage 1 thresholds in the raw-to-raw configuration. For each dataset, the table reports the number and percentage of subjects falling below the caudate threshold, the putamen threshold, and the overall N+ classification under the either-or rule. Datasets are ordered according to the train/test split adopted in the retraining framework.

Dataset	N	Under caudate, n (%)	Under putamen, n (%)	N+, n (%)
LA5c	121	1 (0.83)	3 (2.48)	4 (3.31)
SALD	491	26 (5.30)	10 (2.04)	34 (6.92)
NKI-RS	894	38 (4.25)	37 (4.14)	72 (8.05)
Cam-CAN	627	38 (6.06)	49 (7.81)	73 (11.64)
PREVENT-AD	301	6 (1.99)	2 (0.66)	7 (2.33)
OASIS-3	631	45 (7.13)	51 (8.08)	77 (12.20)
HCP-YoungAdult	1006	77 (7.65)	228 (22.66)	264 (26.24)
HCP-Aging	599	23 (3.84)	49 (8.18)	69 (11.52)
DLBS	414	20 (4.83)	35 (8.45)	47 (11.35)



(a) Application of the retrained raw-to-raw HD-ISS Stage 1 caudate threshold to the raw test datasets.



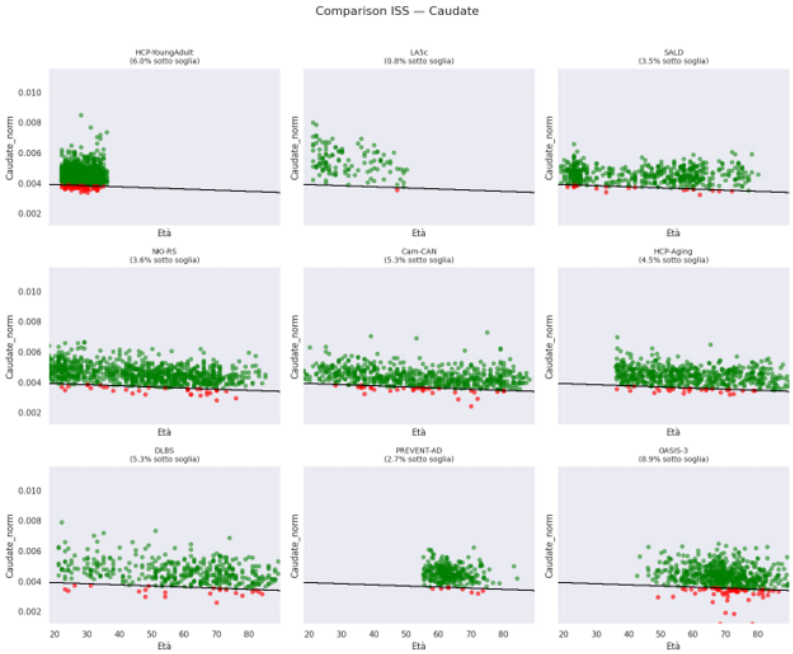
(b) Application of the retrained raw-to-raw HD-ISS Stage 1 putamen threshold to the raw test datasets.

Figure 6.7. Application of the retrained HD-ISS Stage 1 age-dependent thresholds in the raw-to-raw configuration, shown separately for the caudate (top) and putamen (bottom).

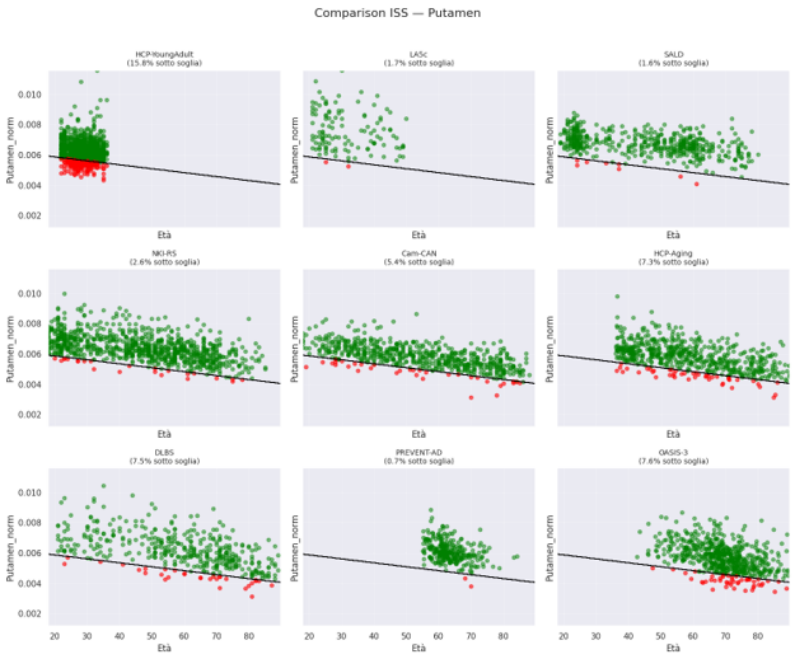
Harmonized-to-raw configuration

Table 6.5. Application of the retrained HD-ISS Stage 1 thresholds in the harmonized-to-raw configuration. For each dataset, the table reports the number and percentage of subjects falling below the caudate threshold, the putamen threshold, and the overall N+ classification under the either-or rule. Datasets are ordered according to the train/test split adopted in the retraining framework.

Dataset	N	Under caudate, n (%)	Under putamen, n (%)	N+, n (%)
LA5c	121	1 (0.83)	2 (1.65)	3 (2.48)
SALD	491	17 (3.46)	8 (1.63)	24 (4.89)
NKI-RS	894	32 (3.58)	23 (2.57)	53 (5.93)
Cam-CAN	627	33 (5.26)	34 (5.42)	55 (8.77)
PREVENT-AD	301	8 (2.66)	2 (0.66)	9 (2.99)
OASIS-3	631	56 (8.87)	48 (7.61)	85 (13.47)
HCP-YoungAdult	1006	60 (5.96)	159 (15.81)	195 (19.38)
HCP-Aging	599	27 (4.51)	44 (7.35)	66 (11.02)
DLBS	414	22 (5.31)	31 (7.49)	44 (10.63)



(a) Application of the retrained harmonized-to-raw HD-ISS Stage 1 caudate threshold to the raw test datasets.



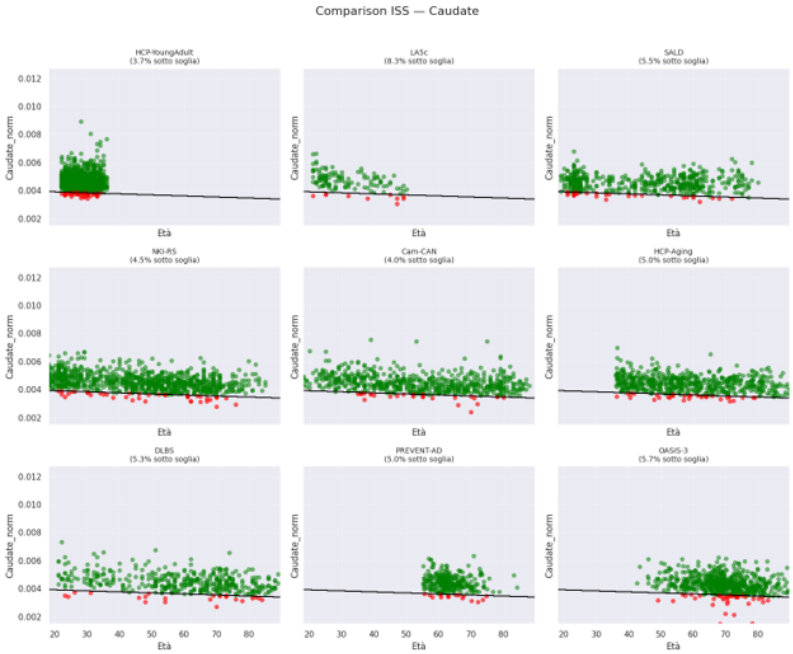
(b) Application of the retrained harmonized-to-raw HD-ISS Stage 1 putamen threshold to the raw test datasets.

Figure 6.8. Application of the retrained HD-ISS Stage 1 age-dependent thresholds in the harmonized-to-raw configuration, shown separately for the caudate (top) and putamen (bottom).

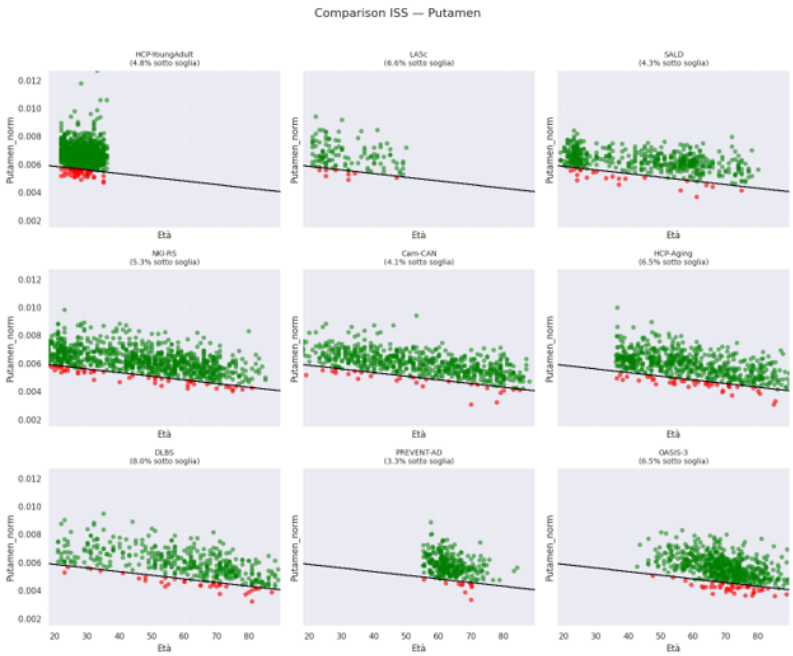
Harmonized-to-harmonized configuration

Table 6.6. Application of the retrained HD-ISS Stage 1 thresholds in the harmonized-to-harmonized configuration. For each dataset, the table reports the number and percentage of subjects falling below the caudate threshold, the putamen threshold, and the overall N+ classification under the either-or rule. Datasets are ordered according to the train/test split adopted in the retraining framework.

Dataset	N	Under caudate, n (%)	Under putamen, n (%)	N+, n (%)
LA5c	121	10 (8.26)	8 (6.61)	14 (11.57)
SALD	491	27 (5.50)	21 (4.28)	43 (8.76)
NKI-RS	894	40 (4.47)	47 (5.26)	82 (9.17)
Cam-CAN	627	25 (3.99)	26 (4.15)	42 (6.70)
PREVENT-AD	301	15 (4.98)	10 (3.32)	22 (7.31)
OASIS-3	631	36 (5.71)	41 (6.50)	64 (10.14)
HCP-YoungAdult	1006	37 (3.68)	48 (4.77)	80 (7.95)
HCP-Aging	599	30 (5.01)	39 (6.51)	64 (10.68)
DLBS	414	22 (5.31)	33 (7.97)	46 (11.11)



(a) Application of the retrained harmonized-to-harmonized HD-ISS Stage 1 caudate threshold to the harmonized test datasets.

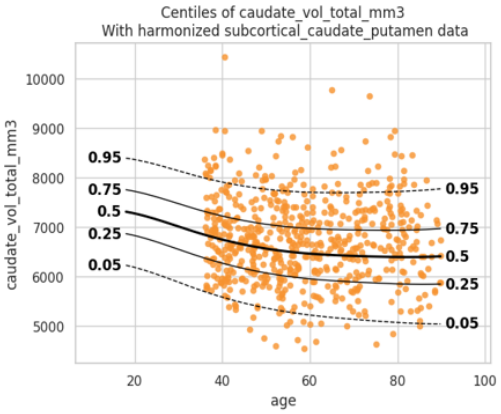


(b) Application of the retrained harmonized-to-harmonized HD-ISS Stage 1 putamen threshold to the harmonized test datasets.

Figure 6.9. Application of the retrained HD-ISS Stage 1 age-dependent thresholds in the harmonized-to-harmonized configuration, shown separately for the caudate (top) and putamen (bottom).

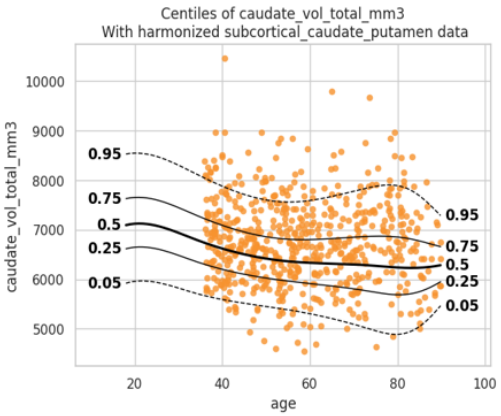
Normative HBR Models

HCP-Aging



(a) HBR Gaussian model, caudate.

(b) HBR Gaussian model, putamen.

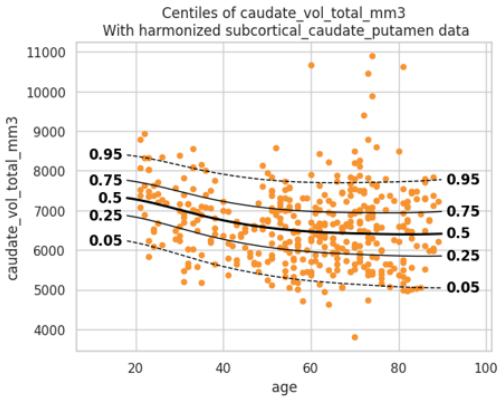


(c) HBR SHASH model, caudate.

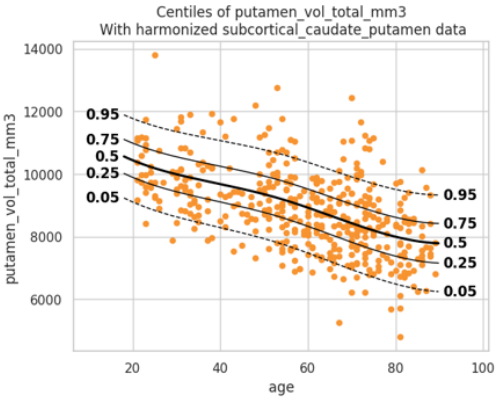
(d) HBR SHASH model, putamen.

Figure 6.10. Estimated centile curves for caudate and putamen volumes in the HCP-Aging dataset obtained using hierarchical Bayesian regression normative models with Gaussian and SHASH likelihoods. The top row shows the HBR Gaussian model and the bottom row the HBR SHASH model; the left column reports the caudate and the right column the putamen. Orange points correspond to observed volumes, solid black lines indicate the estimated 0.25, 0.5, and 0.75 centiles, and dashed lines indicate the 0.05 and 0.95 centiles.

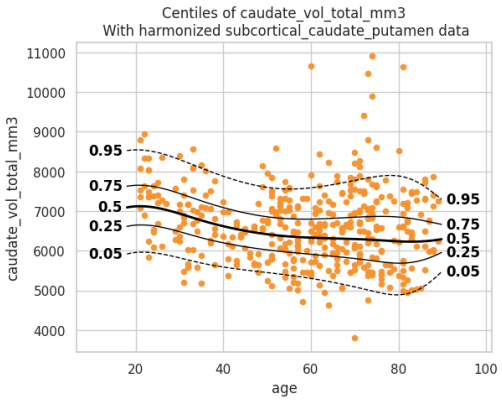
DLBS



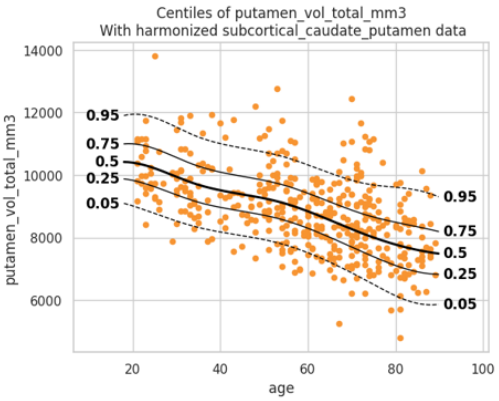
(a) HBR Gaussian model, caudate.



(b) HBR Gaussian model, putamen.



(c) HBR SHASH model, caudate.



(d) HBR SHASH model, putamen.

Figure 6.11. Estimated centile curves for caudate and putamen volumes in the DLBS dataset obtained using hierarchical Bayesian regression normative models with Gaussian and SHASH likelihoods. The top row shows the HBR Gaussian model and the bottom row the HBR SHASH model; the left column reports the caudate and the right column the putamen. Orange points correspond to observed volumes, solid black lines indicate the estimated 0.25, 0.5, and 0.75 centiles, and dashed lines indicate the 0.05 and 0.95 centiles.