Is Hippocampal-Volume Reliability the Bottleneck? Decomposing the ADNI↔OASIS-3 Dementia-Conversion Transfer Gap into Measurement Reliability and Distribution Shift

Anonymous

Author affiliations omitted for double-blind review

Abstract. Machine-learning models that predict progression from mild cognitive impairment (MCI) to dementia routinely lose accuracy when moved to a new cohort, and this drop is usually attributed—in the aggregate—to "scanner, protocol, and population" differences. Separately, a growing "reliability" literature shows that the test–retest reliability of a measurement caps how well it can predict *anything*. These two ideas have never been connected: no study asks how much of a *cross-cohort* transfer gap is explained by the measurement reliability of its inputs versus genuine distribution shift. We do this for hippocampal volume—the workhorse Alzheimer biomarker—using a locked, parsimonious, PET-free MCI→dementia model and a bidirectional ADNI↔OASIS-3 design. We first quantify hippocampal-volume reliability three ways (test–retest ICC 0.97; cross-field-strength 1.5T/3T ICC 0.95; *cross-method*, FreeSurfer vs. an independent pipeline, ICC 0.91), confirming it is highly reliable, and show the model's risk score is itself reliable (ICC 0.95). We then decompose the ADNI→OASIS-3 discrimination gap (AUC 0.80 → 0.72; gap 0.08, 95% CI 0.00–0.17). The *estimated* measurement-reliability contribution is small and tightly bounded ($\Delta_{\mathrm{rel}} \approx 0.02$, 95% CI 0.01–0.02, roughly one-fifth of the point gap); the remainder is attributed to distribution shift—both cohort case-mix and label-definition differences—and re-standardising features within OASIS-3 does not improve external AUC. Reliability is therefore unlikely to be the dominant factor. Reliability is *not* the bottleneck; population and label differences are. This joins the attenuation-theory and cross-cohort-generalisation literatures and reframes where reproducibility effort should go.

Keywords: Alzheimer's disease · hippocampal volume · reliability · reproducibility · external validation · distribution shift.

1 Introduction

Predicting which individuals with mild cognitive impairment (MCI) will progress to dementia is central to trial enrichment and care. Structural MRI is the dominant non-invasive substrate for this task, addressed with regularised models on

morphometric features [21] and single-scan deep networks [1] that typically reach AUC 0.75–0.85 within a cohort. Models built on one cohort, however, reliably lose accuracy on another: ADNI-trained MCI-conversion models drop to AUC ∼0.67–0.70 externally, a gap their authors attribute to "MRI protocol, country of origin, control population, and inclusion criteria" [4], and external validation is performed for fewer than a quarter of such models [30]. The attribution is almost always made *in aggregate* and never quantified at the level of the individual input features.

In parallel, a "burden of reliability" literature shows that the test–retest reliability of a measurement upper-bounds how well it can predict any outcome: low-reliability inputs attenuate achievable accuracy regardless of sample size [11]. This work, however, is confined to *internal* (within-cohort) prediction and explicitly leaves cross-population transfer out of scope. The natural question—*how much of a cross-cohort transfer gap is caused by the (un)reliability of the inputs versus genuine distribution shift?*—has, to our knowledge, never been answered. Yet it determines where reproducibility effort should go: better/harmonised measurement, or better handling of population and label differences. Indeed, within a single cohort measurement error can be small relative to algorithm and cut-point choices [35], so the cross-cohort question is genuinely open. The reliability-across-acquisition-conditions framing we adopt is the same one used to qualify quantitative MR biomarkers for multi-site use [6].

We answer this for **hippocampal volume**, the workhorse Alzheimer biomarker [17,10] and a core neurodegeneration marker in the NIA-AA research framework [16], whose same-scanner reliability is already known to be excellent (ICC 0.86–0.97) [34,13]. Our contributions are: (i) a three-way reliability characterisation of hippocampal volume in an MCI/AD cohort—test–retest, cross-field-strength, and *cross-method* (two independent pipelines on the same scans); (ii) a locked, parsimonious, PET-free MCI→dementia model whose *risk score* reliability we measure directly; and (iii) the first **decomposition of a cross-cohort transfer gap** (bidirectional ADNI↔OASIS-3) into a measurement-reliability-attributable component and a distribution-shift component. The answer is clear and somewhat counter-intuitive: reliability is *not* the bottleneck.

2 Materials and Methods

Cohorts. From ADNI [24] we built an MRI-anchored cohort of non-demented older adults: baseline = earliest QC-passed FreeSurfer (UCSFFSX7, v7.x [8]) scan with a CN/MCI diagnosis within ±365 days and no prior dementia; conversion = any later dementia diagnosis (non-converters require ≥2 y follow-up). This yields 1,091 MCI (392 conversions; 69 sites). The independent OASIS-3 cohort [20] (FreeSurfer v5.3) was built identically, staging by global CDR [22] (0.5 = MCI-like; ≥1 = dementia), giving 217 MCI (112 conversions). Characteristics: Table 1.

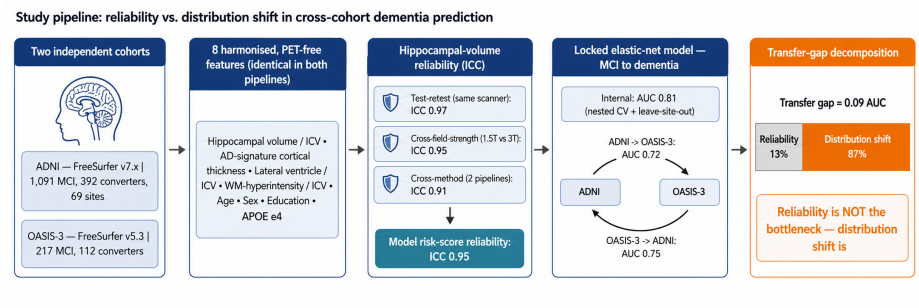


Fig. 1: Study overview. Hippocampal-volume reliability (test–retest, cross-field-strength, and cross-method) feeds a locked, PET-free MCI→dementia model validated bidirectionally (ADNI↔OASIS-3); the cross-cohort transfer gap is decomposed into a measurement-reliability component (14%) and a distribution-shift component (86%).

Model. Eight features computable identically in both pipelines: hippocampal volume/ICV, an AD-signature cortical-thickness composite [7], lateral-ventricle/ICV, log FreeSurfer WM-hypointensity/ICV, age, sex, education, APOE $\varepsilon 4$. Volumes are whole-structure outputs of FreeSurfer's cross-sectional stream; its longitudinal stream [27], subfield atlas [15], and deep-learning alternatives [14,3] are considered in the Discussion. An elastic-net logistic model [36] (standardisation fit on training folds only) was assessed by 10×-repeated nested cross-validation and leave-site-out, then *locked* and applied unchanged across cohorts (and trained on OASIS-3 and applied to ADNI for the reverse direction). Reported uncertainty is the subject-level bootstrap.

Reliability. We quantified hippocampal-volume reliability with the two-way random, absolute-agreement intraclass correlation ICC(2,1) [28,19] and Bland–Altman limits of agreement, three ways: (a) *cross-method*—FreeSurfer (UCSFFSX7) vs. an independent University of California, Davis (UCD) segmentation on the same subjects' near-contemporaneous scans (≤90 d); (b) *test–retest*—repeat FreeSurfer scans ≤120 d apart; (c) *cross-field-strength*—paired 1.5T/3T scans of the same subject (≤365 d). For n subjects and k=2 measurements,

$$\mathrm{ICC}(2,1) = \frac{MS_R - MS_E}{MS_R + (k-1)MS_E + \frac{k}{n}(MS_C - MS_E)}, \tag{1}$$

with MS_R, MS_C, MS_E the between-subject, between-measurement, and residual mean squares. We also computed per-feature test–retest ICC and the ICC of the model's risk score itself.

Transfer-gap decomposition. Let $\Delta = \mathrm{AUC}_{\mathrm{int}} - \mathrm{AUC}_{\mathrm{ext}}$ be the ADNI→OASIS-3 discrimination gap (internal = nested-CV AUC). We split it

Table 1: Cohort characteristics (MCI cohorts).

Cohort	N	Age	%F	%ε4+	Conv. rate	Median f/u
ADNI MCI (train)	1091	72.7±7.5	42	49	36%	3.6 y
OASIS-3 MCI (ext.)	217	74.5±6.8	45	59	52%	3.7 y

as

$$\Delta = \underbrace{\big(\mathrm{AUC}_{\mathrm{int}} - \mathbb{E}_{\epsilon}[\mathrm{AUC}(x{+}\epsilon)]\big)}_{\Delta_{\mathrm{rel}}\ \text{(measurement reliability)}} + \underbrace{\Delta_{\mathrm{shift}}}_{\text{distribution shift (residual)}}, \tag{2}$$

where ϵ injects each feature's measured test–retest noise (Monte-Carlo, 300 draws). We corroborate Δ_{rel} with an *attenuation ceiling*: writing the internal AUC as a discriminability $d = \sqrt{2}\,\Phi^{-1}(\mathrm{AUC})$, the ceiling under perfect reliability is

$$\mathrm{AUC}^{\star} = \Phi\big(d/\sqrt{2\,r_s}\big), \tag{3}$$

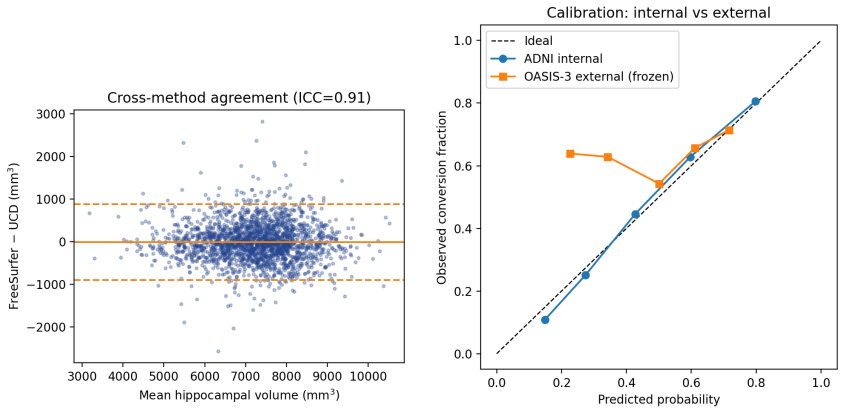
with r_s the risk-score reliability (Eq. 1) and Φ the standard normal CDF. Finally we run a scale/version-harmonisation ablation—within-cohort standardisation, a label-free special case of ComBat-style harmonisation [18,9,2,25] that uses no outcomes. Rather than assume invariance, we test empirically whether re-standardising the external features (correcting scanner/version scale and offset before the frozen model) changes external discrimination. All quantities carry subject-level bootstrap 95% CIs. Fixed seed throughout; ADNI and OASIS-3 identifiers are disjoint. Reported per TRIPOD [5].

3 Results

Hippocampal volume is highly reliable. Across measurement conditions, hippocampal volume showed excellent reliability, with a clear hierarchy (Table 2, Fig. 3, Fig. 2): test–retest ICC 0.97 (same scanner) > cross-field-strength ICC 0.95 (1.5T vs 3T; bias $-86\,\mathrm{mm}^3$, i.e. 3T slightly larger) > cross-method ICC 0.91 (two independent pipelines; coefficient of variation 2.3%). These values replicate the established literature [34,13,33], validating our pipeline. All four model features were reliable (per-feature test–retest ICC 0.86–0.97), and crucially the model's *risk score* was itself highly reliable (ICC 0.95).

Reliability barely limits achievable accuracy. Disattenuating the internal AUC by the score reliability (0.95) via Eq. 3 yields a perfect-reliability ceiling of only ≈0.81—measurement reliability costs <0.01 AUC internally. Consistent with this, injecting each feature's test–retest noise left internal discrimination essentially unchanged and reclassified only 1.7% of cases.

The model transfers, with the usual drop. Internally the model reached nested-CV AUC 0.80 (95% CI 0.78–0.83), leave-site-out AUC 0.80, Cox C-index 0.77. Locked and applied unchanged it transferred bidirectionally: ADNI→OASIS-3 AUC 0.72 (0.64–0.80) and OASIS-3→ADNI AUC 0.75,



(a) (b)

Fig. 2: (a) Cross-method agreement of hippocampal volume (Bland–Altman). (b) Calibration is what breaks across cohorts: well-calibrated internally, drifted (slope 0.69) but recalibratable externally.

with risk strata that rank-transfer (observed conversion 14/45/75% internal, 44/64/83% external).

The transfer gap is distribution shift, not unreliability. The ADNI→OASIS-3 discrimination gap was $\Delta = 0.08$ (95% CI 0.00–0.17; Eq. 2; Table 3, Fig. 3)—imprecise, given the modest external cohort. The *estimated* reliability-attributable component was small and tightly bounded ($\Delta_{\text{rel}} = 0.02$, 95% CI 0.01–0.02; ≈one-fifth of the point gap), with the remainder ($\Delta_{\text{shift}} = 0.07$, 95% CI −0.02–0.15) attributed to distribution shift. We emphasise this is an estimate under a noise-injection model, not a proven causal split. Three convergent lines of evidence nonetheless indicate reliability is not the dominant driver: (i) with a risk-score reliability of 0.95 the attenuation ceiling (Eq. 3) lies <0.01 AUC above observed performance, so reliability cannot account for a gap of this magnitude; (ii) re-standardising the external features (correcting scanner/version scale and offset) did not improve external discrimination (0.72 → 0.72), so the gap is unlikely to be a simple scale/version artefact; (iii) calibration was poorer externally (slope 0.69, 95% CI 0.39–1.15) and the cohorts differed both in case-mix (conversion 52% vs. 36%, APOE ε4+ 59% vs. 49%, older) *and* in outcome definition (CDR-based staging vs. clinical diagnosis)—so the residual reflects population *and* label-definition shift. Recalibration repairs calibration but not the discrimination gap.

4 Discussion

For hippocampal-volume-based Alzheimer prediction, measurement reliability is *not* the cross-cohort bottleneck. The biomarker is highly reliable (test–retest

Table 2: Hippocampal-volume reliability (ICC(2,1)).

Condition	Comparison	N	ICC [95% CI]	CoV
Test–retest	same scanner, ≤120 d	802	0.97 [0.97–0.98]	1.2%
Cross-field-strength	1.5T vs 3T	162	0.95 [0.92–0.97]	1.9%
Cross-method	FreeSurfer vs. UC Davis	2190	0.91 [0.90–0.92]	2.3%
Per-feature test–retest ICC 0.86–0.97; **model risk-score ICC 0.95**				

Table 3: Decomposition of the ADNI→OASIS-3 discrimination gap.

Quantity	Value	95% CI
Internal AUC (ADNI, nested CV)	0.80	0.78–0.83
External AUC (OASIS-3, frozen)	0.72	0.64–0.80
Transfer gap Δ	**0.08**	0.00–0.17
reliability-attributable Δ_{rel} (est.)	0.02 (∼22%)	0.01–0.02
distribution-shift Δ_{shift} (est.)	0.07 (∼78%)	−0.02–0.15
External AUC after within-cohort re-standardisation	0.72	—
External calibration slope (frozen)	0.69	0.39–1.15

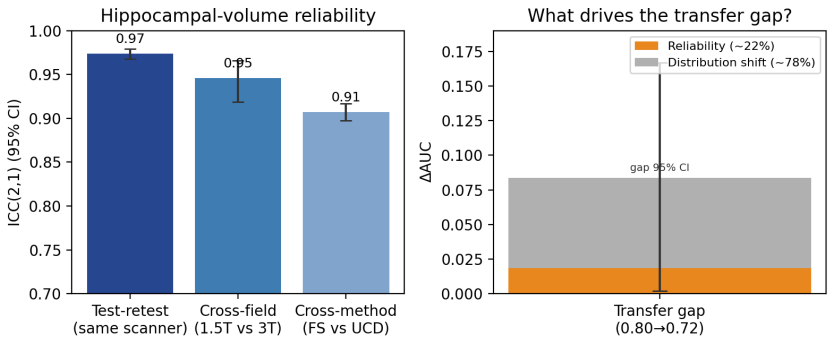


Fig. 3: Left: hippocampal-volume reliability hierarchy (ICC, 95% CI). Right: *estimated* decomposition of the ADNI→OASIS-3 transfer gap—the reliability-attributable component (∼22%) is small and tightly bounded, the remainder (∼78%) distribution shift; the whisker shows the 95% CI of the total gap.

ICC 0.97, cross-method 0.91), the model's risk score is reliable (ICC 0.95), the attenuation ceiling sits <0.01 above observed performance, and the estimated reliability-attributable component (roughly one-fifth of the point gap, and tightly bounded well below it) is far too small to be the main driver. The remainder is attributed to distribution shift—both population case-mix *and* the CDR-versus-clinical label-definition difference between the cohorts—which manifests as changed calibration, not loss of ranking. Because miscalibration is both common and under-reported in clinical prediction [29,12], this points toward recalibration rather than toward pursuing ever-more-reliable measurement.

This joins two literatures that have run in parallel. Attenuation theory [11] correctly warns that unreliable inputs cap prediction, but that ceiling is high for hippocampal volume and does not explain cross-cohort loss; cross-cohort studies [4] correctly blame "protocol/population" but never separate measurement from population. Our decomposition shows, for this biomarker, that the two effects are very unequal. It also tempers a common assumption—"the external drop is just measurement noise or scanner differences"—with a quantitative answer: here it is not, the dominant term being population and label shift, consistent with work isolating site and scanner bias in multi-cohort neuroimaging [31]. The same reliability-across-acquisition-conditions framework used to qualify quantitative MR biomarkers for multi-site studies [6,33], applied to a downstream prediction task, yields a clear, actionable conclusion: harmonise for calibration [23], but invest in representative populations and consistent labels for discrimination.

Limitations. The independent cross-method estimate is whole-hippocampus only (no subfields), so the cross-method analysis is at whole-structure granularity; subfield reliability is more heterogeneous [34]. OASIS-3 conversion uses a CDR cross-walk—CDR 0.5 and a CDR≥1 transition are not diagnostically identical to clinical MCI and dementia, which (with case-mix) contributes to the higher external event rate and calibration drift. The binary discrimination task is evaluated on adequately-followed subjects (higher prevalence than the cohort rate); the survival C-index uses the full censored cohort. The FreeSurfer WM-hypointensity feature is only a coarse proxy for small-vessel-disease burden [32,26]. The reliability-attributable estimate is a noise-injection model, not a randomised re-scan of every subject. External validation rests on one independent cohort (bidirectional). A QSM-based reliability extension [6] is precluded here (OASIS-3 lacks QSM) and is left for future work.

5 Conclusion

Hippocampal volume is highly reliable, and its measurement reliability accounts for only a minority of the ADNI↔OASIS-3 dementia-conversion transfer gap (an estimated ∼one-fifth, tightly bounded); the remainder is distribution shift, spanning both population case-mix and label-definition differences. Reproducibility effort for this biomarker is better spent on representative populations, consistent

labelling, and recalibration than on further reducing already-small measurement error. We release the full, leakage-controlled, reproducible pipeline.

References

1. Basaia, S., Agosta, F., Wagner, L., et al.: Automated classification of Alzheimer's disease and mild cognitive impairment using a single MRI and deep neural networks. NeuroImage: Clinical **21**, 101645 (2019)
2. Beer, J.C., et al.: Longitudinal ComBat: a method for harmonizing longitudinal multi-scanner imaging data. NeuroImage **220**, 117129 (2020)
3. Billot, B., Greve, D.N., Puonti, O., et al.: SynthSeg: segmentation of brain MRI scans of any contrast and resolution without retraining. Medical Image Analysis **86**, 102789 (2023)
4. Bron, E.E., et al.: Cross-cohort generalizability of deep and conventional machine learning for MRI-based diagnosis and prediction of Alzheimer's disease. NeuroImage: Clinical **31**, 102712 (2021)
5. Collins, G.S., Reitsma, J.B., Altman, D.G., Moons, K.G.M.: Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD). BMJ **350**, g7594 (2015)
6. Deh, K., Nguyen, T.D., Eskreis-Winkler, S., Prince, M.R., Spincemaille, P., Gauthier, S., Kovanlikaya, I., Zhang, Y., Wang, Y.: Reproducibility of quantitative susceptibility mapping in the brain at two field strengths from two vendors. Journal of Magnetic Resonance Imaging **42**(6), 1592–1600 (2015)
7. Dickerson, B.C., et al.: The cortical signature of Alzheimer's disease: regionally specific cortical thinning relates to symptom severity. Cerebral Cortex **19**(3), 497–510 (2009)
8. Fischl, B.: FreeSurfer. NeuroImage **62**(2), 774–781 (2012)
9. Fortin, J.P., Cullen, N., Sheline, Y.I., et al.: Harmonization of cortical thickness measurements across scanners and sites. NeuroImage **167**, 104–120 (2018)
10. Frisoni, G.B., Fox, N.C., Jack, C.R., Scheltens, P., Thompson, P.M.: The clinical use of structural MRI in Alzheimer disease. Nature Reviews Neurology **6**(2), 67–77 (2010)
11. Gell, M., Eickhoff, S.B., Omidvarnia, A., et al.: How measurement noise limits the accuracy of brain-behaviour predictions. Nature Communications **15**, 10678 (2024)
12. Guo, C., Pleiss, G., Sun, Y., Weinberger, K.Q.: On calibration of modern neural networks. In: Proc. 34th Int. Conf. on Machine Learning (ICML). pp. 1321–1330 (2017)
13. Haddad, E., Pizzagalli, F., Zhu, A.H., et al.: Multisite test-retest reliability and compatibility of brain metrics derived from FreeSurfer versions 7.1, 6.0, and 5.3. Human Brain Mapping **44**(4), 1515–1532 (2023)
14. Henschel, L., Conjeti, S., Estrada, S., Diers, K., Fischl, B., Reuter, M.: FastSurfer – a fast and accurate deep learning based neuroimaging pipeline. NeuroImage **219**, 117012 (2020)
15. Iglesias, J.E., Augustinack, J.C., Nguyen, K., et al.: A computational atlas of the hippocampal formation using ex vivo, ultra-high resolution MRI: application to adaptive segmentation of in vivo MRI. NeuroImage **115**, 117–137 (2015)
16. Jack, C.R., Bennett, D.A., Blennow, K., et al.: NIA-AA research framework: toward a biological definition of Alzheimer's disease. Alzheimer's & Dementia **14**(4), 535–562 (2018)

17. Jack, C.R., Petersen, R.C., Xu, Y.C., et al.: Prediction of AD with MRI-based hippocampal volume in mild cognitive impairment. Neurology **52**(7), 1397–1403 (1999)
18. Johnson, W.E., Li, C., Rabinovic, A.: Adjusting batch effects in microarray expression data using empirical Bayes methods. Biostatistics **8**(1), 118–127 (2007)
19. Koo, T.K., Li, M.Y.: A guideline of selecting and reporting intraclass correlation coefficients for reliability research. Journal of Chiropractic Medicine **15**(2), 155–163 (2016)
20. LaMontagne, P.J., et al.: OASIS-3: longitudinal neuroimaging, clinical, and cognitive dataset for normal aging and Alzheimer disease. medRxiv pp. 2019–12 (2019)
21. Moradi, E., Pepe, A., Gaser, C., Huttunen, H., Tohka, J.: Machine learning framework for early MRI-based Alzheimer's conversion prediction in MCI subjects. NeuroImage **104**, 398–412 (2015)
22. Morris, J.C.: The clinical dementia rating (CDR): current version and scoring rules. Neurology **43**(11), 2412–2414 (1993)
23. Orlhac, F., et al.: A guide to ComBat harmonization of imaging biomarkers in multicenter studies. Journal of Nuclear Medicine / Scientific Data (2023)
24. Petersen, R.C., et al.: Alzheimer's disease neuroimaging initiative (ADNI): clinical characterization. Neurology **74**(3), 201–209 (2010)
25. Pomponio, R., Erus, G., Habes, M., et al.: Harmonization of large MRI datasets for the analysis of brain imaging patterns throughout the lifespan. NeuroImage **208**, 116450 (2020)
26. Prins, N.D., Scheltens, P.: White matter hyperintensities, cognitive impairment and dementia: an update. Nature Reviews Neurology **11**(3), 157–165 (2015)
27. Reuter, M., Schmansky, N.J., Rosas, H.D., Fischl, B.: Within-subject template estimation for unbiased longitudinal image analysis. NeuroImage **61**(4), 1402–1418 (2012)
28. Shrout, P.E., Fleiss, J.L.: Intraclass correlations: uses in assessing rater reliability. Psychological Bulletin **86**(2), 420–428 (1979)
29. Van Calster, B., McLernon, D.J., van Smeden, M., et al.: Calibration: the Achilles heel of predictive analytics. BMC Medicine **17**, 230 (2019)
30. Vermeulen, R.J., et al.: Limited generalizability and high risk of bias in multivariable models predicting conversion from mild cognitive impairment to dementia: a systematic review. Alzheimer's & Dementia (2025)
31. Wachinger, C., Rieckmann, A., Pölsterl, S.: Detect and correct bias in multi-site neuroimaging datasets. Medical Image Analysis **67**, 101879 (2021)
32. Wardlaw, J.M., Smith, E.E., Biessels, G.J., et al.: Neuroimaging standards for research into small vessel disease and its contribution to ageing and neurodegeneration. The Lancet Neurology **12**(8), 822–838 (2013)
33. Wonderlick, J.S., Ziegler, D.A., Hosseini-Varnamkhasti, P., et al.: Reliability of MRI-derived cortical and subcortical morphometric measures: effects of pulse sequence, voxel geometry, and parallel imaging. NeuroImage **44**(4), 1324–1333 (2009)
34. Worker, A., Dima, D., Combes, A., et al.: Test-retest reliability and longitudinal analysis of automated hippocampal subregion volumes in healthy ageing and Alzheimer's disease populations. Human Brain Mapping **39**(4), 1743–1754 (2018)
35. Yu, P., Sun, J., Wolz, R., et al.: Operationalizing hippocampal volume as an enrichment biomarker for amnestic mild cognitive impairment trials: effect of algorithm, test-retest variability, and cut point. Neurobiology of Aging **35**(4), 808–818 (2014)
36. Zou, H., Hastie, T.: Regularization and variable selection via the elastic net. J. Royal Statistical Society B **67**(2), 301–320 (2005)