

Auditing missingness and input-outcome circularity in unsupervised clinical phenotyping: an empirical case study in hospitalized traumatic brain injury

Likun Yang

The 904th Hospital of the Joint Logistics Support Force of the People's Liberation Army

Wei Lin

lin.wei@mayo.edu

The 904th Hospital of the Joint Logistics Support Force of the People's Liberation Army

Yan Wu

The 904th Hospital of the Joint Logistics Support Force of the People's Liberation Army

Liang Zhang

The 904th Hospital of the Joint Logistics Support Force of the People's Liberation Army

Dan Wang

The 904th Hospital of the Joint Logistics Support Force of the People's Liberation Army

Yuhai Wang

The 904th Hospital of the Joint Logistics Support Force of the People's Liberation Army

Research Article

Keywords: missing data, unsupervised learning, cluster analysis, clinical phenotyping, circular analysis, multiple imputation, traumatic brain injury, measurement methodology

Posted Date: August 10th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-10251325/v1>

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

[Read Full License](#)

Additional Declarations: No competing interests reported.

Abstract

****Background:**** Routinely collected clinical scale data are increasingly used for unsupervised phenotyping, but two design choices are often under-audited: how missing scale values are handled, and whether variables later interpreted as outcomes were also used to define the clusters. We used a hospitalized traumatic brain injury (TBI) scale dataset as an empirical case study to audit missingness sensitivity and input-outcome circularity in unsupervised clinical phenotyping.

****Methods:**** We analysed 582 single-centre hospitalized TBI records with five routinely collected scales: Mini-Mental State Examination (MMSE), Montreal Cognitive Assessment (MoCA), 15-item Geriatric Depression Scale (GDS), 17-item Hamilton Depression Rating Scale, and Barthel Activities of Daily Living (ADL) Index. The reference pipeline used median imputation, z-standardisation, and Gaussian mixture modelling (GMM), with component number selected by Bayesian information criterion. The methodological endpoints were cluster-wise bootstrap stability, cross-method partition concordance, sensitivity to multiple imputation by chained equations (MICE, $m=20$), complete-case re-clustering, and change in apparent cluster-function association after removing Barthel ADL from the clustering input.

****Results:**** Complete five-scale data were available for 103/582 patients (17.7%); Hamilton and Barthel ADL were missing in 216/582 (37.1%) and 171/582 (29.4%) patients, respectively. The reference GMM selected $k=5$, but cluster-wise bootstrap Jaccard coefficients were 0.211-0.454 and cross-method consensus adjusted Rand index (ARI) was 0.229. Across 20 MICE datasets, mean ARI against the reference partition was 0.157 (empirical imputation-distribution interval 0.139-0.176). Complete-case re-clustering favoured $k=3$ rather than $k=5$, and the previously emphasised Hamilton-extreme phenotype was adjudicated as an artefact by two of three pre-specified checks. Apparent cluster-function association collapsed when Barthel ADL was removed from the clustering definition: McFadden pseudo- R^2 for three-level ADL severity fell from 0.395 with the original five-variable clusters to 0.005 with four-variable cognitive-affective clusters.

****Conclusions:**** In this empirical case study, a plausible unsupervised clinical phenotype was highly sensitive to missing-data handling and an apparently strong functional association was largely explained by outcome-containing clustering. Unsupervised phenotyping studies using routine clinical scale data should report, at minimum, missingness patterns, cluster stability, imputation sensitivity, complete-case sensitivity when feasible, and input-outcome circularity checks before assigning clinical meaning to derived subtypes.

Background

Routinely collected clinical scale data are attractive inputs for data-driven phenotyping. They are inexpensive, often available at scale, and can capture dimensions that are difficult to summarise in a single diagnosis code. In neurorehabilitation and other hospital settings, cognitive, affective, and functional scales are frequently combined to describe patient state, stratify services, or generate

hypotheses about clinically meaningful subgroups [1]. Unsupervised clustering of such scale batteries can therefore appear to offer a direct route from routine measurement to phenotypic taxonomy [2].

That route has two methodological vulnerabilities. First, routine clinical scales are rarely complete. If missingness is concentrated in particular instruments, the resulting cluster solution may reflect the imputation strategy as much as the underlying patient structure. Multiple imputation can reduce bias when assumptions are defensible, but it should not be treated as a guarantee that an unsupervised partition is stable [3–6]. Second, a variable can be used twice: once as an input defining clusters, and again as a downstream outcome used to validate or interpret those clusters. This input-outcome circularity can produce an apparent external association even when the cluster label adds little information beyond the outcome-like variable already embedded in the clustering definition. The problem is conceptually related to circular analysis or double dipping, where a data-dependent selection step contaminates a subsequent interpretation of the same signal [7].

Hospitalized traumatic brain injury (TBI) provides a useful empirical setting for auditing these issues. Routine TBI assessment may include cognitive screens, mood scales, and functional independence measures [1, 8–11], but these instruments measure partially distinct constructs and are not equally complete in retrospective records. A cluster solution derived from the combined battery may therefore look clinically interpretable while being unstable under alternative missing-data handling or inflated by inclusion of functional scores in the clustering input.

We used a 582-patient hospitalized TBI scale dataset as a methodological case study. The objective was not to propose or validate a TBI phenotype taxonomy. Instead, we operationalised a five-part audit for unsupervised clinical phenotyping: missingness mapping, bootstrap cluster stability, cross-method partition concordance, imputation sensitivity, and input-outcome circularity stress testing. We asked whether a conventional unsupervised phenotyping pipeline would remain stable under resampling, alternative imputation, and complete-case restriction, and whether an apparent cluster-function association would survive removal of the functional variable from the clustering definition. Figure 1 summarises the audit structure.

Methods

Study design and reporting

This was a single-centre retrospective cross-sectional methodological audit of routinely collected clinical scale data. The empirical dataset was used to stress-test an unsupervised phenotyping workflow rather than to estimate causal effects or build a clinical prediction model. Reporting followed STROBE guidance for observational studies [12], with selected TRIPOD + AI items used only where they were relevant to transparent reporting of model-derived claims [13]. The study was approved by the Ethics Committee of The 904th Hospital of the Joint Logistics Support Force of the People's Liberation Army

(approval number 2016002). Informed consent was waived under the institutional policy applicable to fully de-identified retrospective inpatient records.

The primary methodological endpoints were: (1) cluster-wise stability under bootstrap resampling; (2) concordance of the reference partition with alternative clustering methods; (3) concordance of the reference partition with partitions obtained after MICE; (4) persistence of the partition and key phenotype in a complete-case subset; and (5) change in apparent cluster-function association after removing Barthel ADL from the clustering input.

Empirical dataset and scale variables

The dataset comprised 582 patients admitted to the Department of Neurosurgery of a tertiary general hospital in eastern China for whom at least four of five core scales were available. The five scales were MMSE, MoCA, 15-item GDS, 17-item Hamilton Depression Rating Scale, and Barthel ADL Index [8–10]. Selection was determined by scale availability, not by injury severity or rehabilitation-specific eligibility. Glasgow Coma Scale, time since injury, injury mechanism, rater identity, and patient sex were not reliably recoverable from the historical record. The cohort should therefore be interpreted as an acute neurosurgical inpatient TBI sample with incomplete severity characterisation.

Original paper-based forms were digitised with RapidOCR, followed by a 10% random manual spot-check of extracted scale totals. Only total scale scores were used. MoCA subdomain scores and unreliable demographic fields were excluded before analysis. Table 1 summarises the scale availability pattern.

Missing-data assessment

Complete five-scale data were present in 103 patients (17.7%). Item-level missingness was 7.0% for MMSE, 8.8% for MoCA, 0% for GDS, 37.1% for Hamilton, and 29.4% for Barthel ADL. For Hamilton and Barthel ADL, we fitted logistic regressions of the missingness indicator on the other observed scale totals. These observed-scale diagnostics did not provide evidence against a missing-at-random working assumption, but unmeasured predictors of missingness could not be excluded because age, sex, TBI severity, and ward-process variables were unavailable. We therefore interpret MICE analyses as sensitivity analyses rather than definitive correction for missing data.

Reference unsupervised phenotyping pipeline

The reference pipeline used median imputation for missing scale totals, z-standardisation, and Gaussian mixture modelling with full covariance [2]. The number of components was selected by Bayesian information criterion over $k = 2$ to $k = 7$, with a fixed random seed for reproducibility. Cluster quality and concordance were evaluated with silhouette, Calinski-Harabasz, Davies-Bouldin, and partition-concordance metrics, but clinical interpretation was intentionally subordinated to stability and sensitivity results.

Cluster-wise bootstrap stability was assessed with the Jaccard coefficient over 1000 resamples per cluster. Following the cluster-stability literature, values below 0.75 were treated as insufficient for a stable cluster-level interpretation [14]. Cross-method consensus was evaluated by comparing the GMM solution with k-means, Ward agglomerative clustering, and spectral clustering at the same k, summarised by adjusted Rand index [15].

Multiple-imputation sensitivity analysis

The five-scale GMM pipeline was re-run on 20 MICE datasets generated with chained equations using Bayesian ridge regression and posterior sampling. In each imputed dataset, the GMM was fit at the reference $k = 5$ to isolate the effect of imputed values on partition assignment. ARI against the original median-imputed reference partition was calculated for each imputation. The imputation-specific ARI distribution was summarised by its mean, standard deviation, range, and empirical 95% interval across the 20 imputations. Because ARI is a bounded partition-concordance statistic rather than a model coefficient with a conventional within-imputation variance estimate, this interval is treated as a descriptive imputation-distribution interval rather than a formal Rubin-rules inferential confidence interval. Pre-specified interpretation thresholds were $\text{ARI} \geq 0.70$ as robust, 0.40 to < 0.70 as moderately sensitive, and < 0.40 as substantially imputation-dependent.

Complete-case re-clustering

GMM was refit on the 103 patients with complete five-scale data, with BIC allowed to re-select k. We then evaluated whether the Hamilton-extreme cluster highlighted in the reference solution persisted using three pre-specified checks: label coverage of original patients entering the complete-case subset, phenotype-based search for high Hamilton plus low MMSE plus preserved MoCA, and raw-distribution standard-deviation separation in the expected direction. A two-of-three rule determined whether the phenotype was classified as replicating or artefactual.

Input-outcome circularity analysis

To test circularity, Barthel ADL was removed from the clustering input and the GMM was refit on the four cognitive-affective scales (MMSE, MoCA, GDS, and Hamilton). We then modelled three-level ADL severity as the dependent variable in ordinal models using four alternatives as predictors: original five-variable cluster, four-variable cluster, ADL quartile alone, and MMSE quartile alone. Goodness of fit was summarised by McFadden pseudo-R² [16]. A substantial reduction in pseudo-R² after ADL removal was interpreted as evidence that the original cluster-function association was driven by outcome-containing clustering rather than independent latent stratification.

Interpretability probe

A classification and regression tree was fitted as a post hoc interpretability probe [17] to approximate the GMM partition from raw scale scores. It was not treated as a prediction model, clinical instrument, or bedside decision rule; therefore calibration and decision-curve analysis were not applicable.

Software and reproducibility

Analyses were implemented in Python with scikit-learn and standard scientific-computing libraries. Random seeds, MICE settings, and main configuration choices are reported in the supplementary methods. Patient-level data cannot be publicly deposited because institutional policy restricts sharing of retrospective hospital records. The analysis code, derived non-identifiable summary tables, and a synthetic demonstration template can be shared by the corresponding author subject to institutional review; these materials are intended to support reproducibility of the workflow without releasing patient-level records.

Results

Scale availability and missingness

Complete five-scale data were available for 103/582 patients. Missingness was concentrated in Hamilton and Barthel ADL rather than distributed evenly across the battery (Table 1). Logistic missingness diagnostics using observed scale totals were non-significant for the two most affected variables, but the absence of severity and demographic covariates prevented a strong claim that missingness was ignorable.

Reference clustering selected five components but was stability-limited

The reference median-imputed five-scale GMM selected $k = 5$ by BIC. The five clusters included one Hamilton-extreme group ($n = 80$), two minimal-mood intermediate-cognition groups, one moderate-mood lower-MMSE group, and one minimal-mood lower-MoCA group. Despite apparent separation in the fitted model, cluster-wise bootstrap Jaccard coefficients ranged from 0.211 to 0.454, all below the 0.75 stability threshold. Cross-method consensus between GMM and k-means, Ward agglomerative, and spectral clustering was low, with mean pairwise ARI of 0.229 (Table 2).

The partition was highly sensitive to multiple imputation

Across 20 MICE datasets, the mean ARI against the median-imputed reference partition was 0.157 (empirical imputation-distribution interval 0.139–0.176), with per-imputation ARI ranging from 0.088 to 0.235 (Fig. 2). This was well below the pre-specified 0.40 threshold for substantial imputation dependence. The result indicates that the five-cluster reference solution should be interpreted as hypothesis-generating only rather than as a reproducible phenotype taxonomy.

Complete-case re-clustering changed the component structure

In the complete-case subset ($n = 103$), BIC favoured $k = 3$ rather than $k = 5$. The Hamilton-extreme phenotype did not survive the pre-specified adjudication. Although 12 of 14 complete-case survivors from the original Hamilton-extreme cluster co-clustered in one complete-case component, the phenotype-based search met only two of four criteria and raw-distribution separation reversed for MMSE: the complete-case high-Hamilton group had higher, not lower, MMSE than the rest of the subset. The composite verdict was CLUSTER_ARTIFACT by two of three methods.

Apparent functional association collapsed after removing ADL from the input

The input-outcome circularity analysis produced the largest methodological signal. With Barthel ADL included in the clustering input, the original five-variable cluster label was strongly associated with three-level ADL severity (McFadden pseudo- $R^2 = 0.395$). After Barthel ADL was removed and clustering was repeated on MMSE, MoCA, GDS, and Hamilton only, the corresponding pseudo- R^2 fell to 0.005. ADL quartiles alone yielded pseudo- $R^2 = 0.714$, whereas MMSE quartiles alone yielded pseudo- $R^2 = 0.003$. Thus, the functional signal attributed to the original cluster label was largely explained by including the functional score in the cluster definition.

Interpretability probe

The post hoc tree approximation reproduced the four-variable GMM partition with mean five-fold cross-validated accuracy of 0.854 (SD 0.032; largest-class baseline 0.351). Splits used Hamilton, GDS, MoCA, and MMSE. This finding was used only to understand which observed scales drove the partition; it does not constitute validation of a deployable classifier.

Discussion

This empirical audit shows how a plausible unsupervised clinical phenotype can be shaped by missing-data handling and input-outcome circularity. Its main contribution is a concrete five-part stress-test framework for routine clinical scale datasets: mapping missingness, quantifying cluster-wise stability, checking cross-method concordance, testing imputation sensitivity, and auditing overlap between clustering inputs and validation outcomes. In the reference analysis, BIC selected a five-component GMM and individual assignments appeared confident. However, cluster-wise bootstrap stability was low, cross-method concordance was weak, MICE produced partitions poorly aligned with the median-imputed reference solution, and complete-case re-clustering favoured a different component structure. Most importantly, an apparently strong cluster-function association nearly disappeared when the functional variable was removed from the clustering input.

The central methodological lesson is that internal fit and clinical interpretability are not sufficient safeguards for unsupervised clinical phenotyping. Model-based clustering can identify a statistically preferred partition while cluster boundaries remain unstable under resampling or alternative preprocessing [2, 14]. A cluster label can look meaningful because it is partly defined by a clinically salient variable, and it can look reproducible within one imputed dataset while being unstable across plausible alternatives. These problems are especially relevant for routine clinical scale data, where missingness often reflects ward workflow, patient condition, rater availability, or historical documentation rather than a random process.

The missing-data results should be interpreted conservatively. The observed-scale diagnostics did not identify strong predictors of Hamilton or Barthel ADL missingness, but the most clinically important missingness predictors were unavailable. The MICE results therefore do not prove that a particular alternative partition is correct. Rather, they show that the substantive five-cluster interpretation depends heavily on how missing values are filled. In a methods-oriented reading, that dependence is the result.

The ADL-removal analysis provides a direct audit for circular validation, extending the general concern about circular analysis to the specific setting of outcome-containing clinical clustering [7]. Including Barthel ADL in the cluster definition and then using ADL severity to support the clinical meaning of the cluster is an outcome-containing analysis. The collapse in pseudo-R² after ADL removal indicates that the original cluster-function association should not be interpreted as evidence that the cluster label independently stratifies functional status. For clinical researchers using scale batteries, the practical implication is straightforward: if a variable or close surrogate is used to define clusters, later associations with that same construct should be labelled as descriptive internal structure, not external validation.

This case study suggests a minimal reporting package for unsupervised phenotyping with routine clinical scale data. Authors should report variable-level missingness, complete-case size, the imputation model and its assumptions, cluster-wise stability rather than only global fit, concordance under at least one alternative clustering or missing-data strategy, and a circularity audit when downstream outcomes overlap with clustering inputs. When patient-level data cannot be shared, derived summary tables, analysis code, and synthetic or masked examples can still improve reproducibility of the workflow.

The study has limitations. It uses a single-centre acute neurosurgical TBI sample, and the findings are not a validated statement about TBI phenotypes in other care settings. GCS, injury mechanism, time since injury, sex, rater identity, length of stay, ICU admission, and CT lesion type were not reliably recoverable, preventing severity-adjusted interpretation and limiting missingness diagnostics. MMSE and MoCA were the available cognitive screens, while other instruments may be preferable for serial post-traumatic amnesia assessment [18, 19]. Barthel ADL was available as the functional measure, but other TBI outcome measures could behave differently. Finally, the CART analysis was an interpretability probe only and should not be used as a clinical rule.

Conclusions

In this hospitalized TBI scale dataset, unsupervised clinical phenotyping was sensitive to missing-data handling, unstable under resampling, and vulnerable to input-outcome circularity. The case study supports a conservative reporting standard: before translating scale-derived clusters into clinical subtypes, researchers should audit missingness, cluster stability, imputation sensitivity, complete-case behaviour when possible, and overlap between clustering inputs and validation outcomes.

Abbreviations

ADL, Activities of Daily Living; ARI, adjusted Rand index; BIC, Bayesian information criterion; CART, classification and regression tree; GDS, Geriatric Depression Scale; GMM, Gaussian mixture model; MAR, missing at random; MICE, multiple imputation by chained equations; MMSE, Mini-Mental State Examination; MoCA, Montreal Cognitive Assessment; TBI, traumatic brain injury.

Declarations

Ethics approval and consent to participate

This study was approved by the Ethics Committee of The 904th Hospital of the Joint Logistics Support Force of the People's Liberation Army (approval number 2016002). Informed consent was waived under the institutional policy applicable to fully de-identified retrospective inpatient records.

Consent for publication

Not applicable.

Competing interests

The authors declare that they have no competing interests.

Use of AI and automated tools

RapidOCR was used for source-data extraction from paper-based clinical scale forms with 10% manual spot-checking. A large language model was used by the corresponding author for statistical code drafting, with line-by-line human review and independent re-execution, and for English language editing. No AI system is listed as an author, and the corresponding author takes full responsibility for content integrity.

Funding

This research received no external funding.

Author Contribution

W.L. conceived the methodological audit, designed the analysis plan, wrote the analysis code, performed the statistical analyses, prepared the figures and tables, and drafted the manuscript. L.Y., Yan W., L.Z., and D.W. contributed to source-record review, data extraction, manual verification of scale totals, and clinical data curation. Yuhai W. supervised the project, contributed to clinical interpretation, and reviewed the manuscript critically. All authors reviewed and approved the final manuscript.

Acknowledgements

We thank the patients, families, and clinical staff of The 904th Hospital of the Joint Logistics Support Force of the People's Liberation Army. We gratefully acknowledge the contribution of the late Dr Sanjeet S Grewal of Mayo Clinic Florida, whose mentorship shaped the broader research programme from which this analysis emerged.

Availability of data and materials

Patient-level data cannot be publicly deposited because institutional policy restricts sharing of retrospective hospital records. Analysis code, derived non-identifiable summary tables, and a synthetic demonstration template are available from the corresponding author on reasonable request and subject to institutional review.

References

1. Wilde EA, Whiteneck GG, Bogner J, Bushnik T, Cifu DX, Dikmen S, et al. Recommendations for the use of common outcome measures in traumatic brain injury research. *Arch Phys Med Rehabil*. 2010;91:1650–e166017. 10.1016/j.apmr.2010.06.033.
2. Fraley C, Raftery AE. Model-based clustering, discriminant analysis, and density estimation. *J Am Stat Assoc*. 2002;97:611–31. 10.1198/016214502760047131.
3. Sterne JAC, White IR, Carlin JB, Spratt M, Royston P, Kenward MG, Wood AM, Carpenter JR. Multiple imputation for missing data in epidemiological and clinical research: potential and pitfalls. *BMJ*. 2009;338:b2393. 10.1136/bmj.b2393.
4. White IR, Royston P, Wood AM. Multiple imputation using chained equations: issues and guidance for practice. *Stat Med*. 2011;30:377–99. 10.1002/sim.4067.
5. Rubin DB. *Multiple Imputation for Nonresponse in Surveys*. New York: Wiley; 1987. 10.1002/9780470316696.

6. van Buuren S. Flexible Imputation of Missing Data. 2nd ed. Boca Raton: Chapman and Hall/CRC; 2018. 10.1201/9780429492259.
7. Kriegeskorte N, Simmons WK, Bellgowan PSF, Baker CI. Circular analysis in systems neuroscience: the dangers of double dipping. *Nat Neurosci*. 2009;12:535–40. 10.1038/nn.2303.
8. Folstein MF, Folstein SE, McHugh PR. Mini-mental state: a practical method for grading the cognitive state of patients for the clinician. *J Psychiatr Res*. 1975;12:189–98. 10.1016/0022-3956(75)90026-6.
9. Nasreddine ZS, Phillips NA, Bedirian V, Charbonneau S, Whitehead V, Collin I, et al. The Montreal Cognitive Assessment, MoCA: a brief screening tool for mild cognitive impairment. *J Am Geriatr Soc*. 2005;53:695–9. 10.1111/j.1532-5415.2005.53221.x.
10. Yesavage JA, Brink TL, Rose TL, Lum O, Huang V, Adey M, et al. Development and validation of a geriatric depression screening scale: a preliminary report. *J Psychiatr Res*. 1982;17:37–49. 10.1016/0022-3956(82)90033-4.
11. de Guise E, Leblanc J, Champoux MC, Couturier C, Alturki AY, Lamoureux J, et al. The Mini-Mental State Examination and the Montreal Cognitive Assessment after traumatic brain injury: an early predictive study. *Brain Inj*. 2013;27:1428–34. 10.3109/02699052.2013.835868.
12. von Elm E, Altman DG, Egger M, Pocock SJ, Gøtzsche PC, Vandenbroucke JP, STROBE Initiative. The Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) statement: guidelines for reporting observational studies. *PLoS Med*. 2007;4(10):e296. 10.1371/journal.pmed.0040296.
13. Collins GS, Moons KGM, Dhiman P, Riley RD, Beam AL, Van Calster B, et al. TRIPOD + AI statement: updated guidance for reporting clinical prediction models that use regression or machine learning methods. *BMJ*. 2024;385:e078378. 10.1136/bmj-2023-078378.
14. Hennig C. Cluster-wise assessment of cluster stability. *Comput Stat Data Anal*. 2007;52:258–71. 10.1016/j.csda.2006.11.025.
15. Hubert L, Arabie P. Comparing partitions. *J Classif*. 1985;2:193–218. 10.1007/BF01908075.
16. McFadden D. Conditional logit analysis of qualitative choice behavior. In: Zarembka P, editor. *Frontiers in Econometrics*. New York: Academic; 1974. pp. 105–42.
17. Breiman L, Friedman JH, Olshen RA, Stone CJ. *Classification and Regression Trees*. Belmont: Wadsworth; 1984.
18. Levin HS, O'Donnell VM, Grossman RG. The Galveston Orientation and Amnesia Test: a practical scale to assess cognition after head injury. *J Nerv Ment Dis*. 1979;167:675–84. 10.1097/00005053-197911000-00004.
19. Jackson WT, Novack TA, Dowler RN. Effective serial measurement of cognitive orientation in rehabilitation: the Orientation Log. *Arch Phys Med Rehabil*. 1998;79:718–20. 10.1016/S0003-9993(98)90051-X.

Tables

****Table 1. Scale availability in the empirical dataset.****

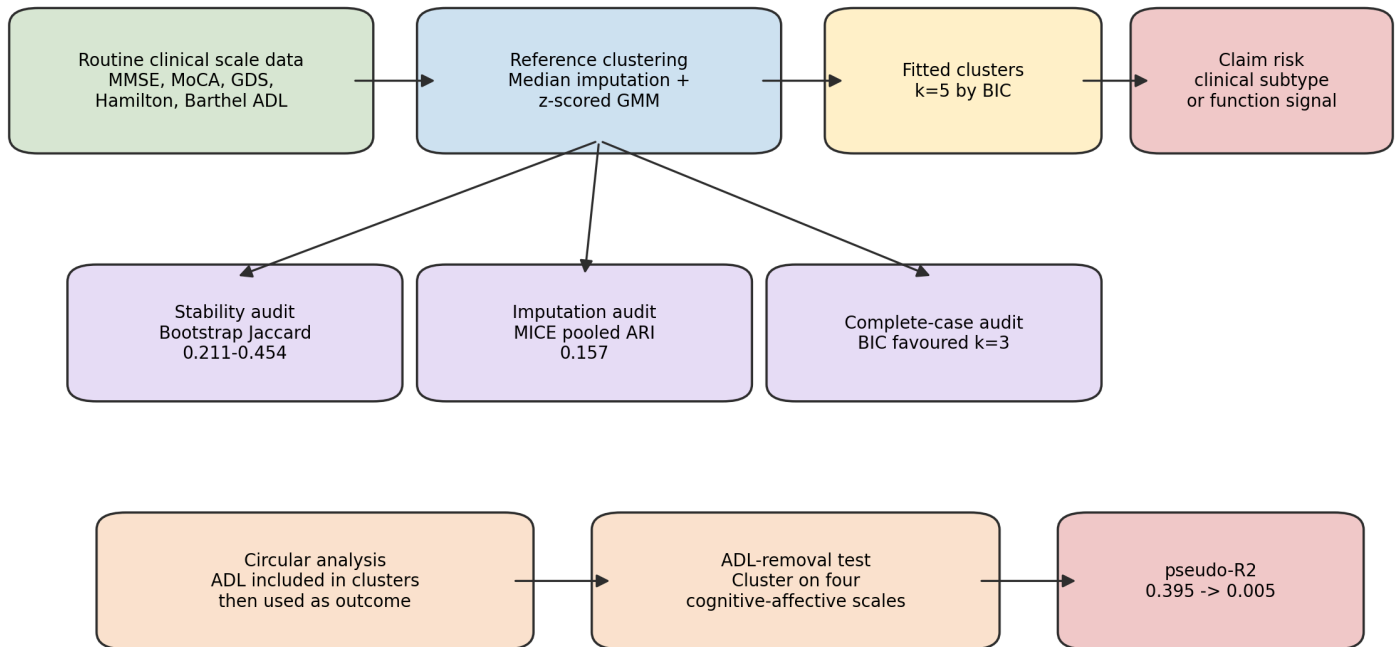
Variable or summary	n/N	% of cohort	Role in reference analysis
MMSE missing	41/582	7.0	Cognitive input
MoCA missing	51/582	8.8	Cognitive input
GDS-15 missing	0/582	0.0	Affective input
Hamilton Depression Rating Scale missing	216/582	37.1	Affective input
Barthel ADL Index missing	171/582	29.4	Functional input and circularity-audit target
Complete five-scale cases	103/582	17.7	Complete-case re-clustering subset

****Table 2. Main methodological robustness findings.****

Audit question	Result	Interpretation
Reference GMM component number	k=5 by BIC	Fitted model preferred a finer partition
Cluster-wise bootstrap Jaccard	0.211-0.454	Below stability threshold
Cross-method consensus ARI	0.229	Low agreement across clustering methods
MICE pooled ARI vs reference	0.157 (95% CI 0.139-0.176)	Substantially imputation-dependent
Complete-case BIC-selected k	k=3	Component structure changed
Hamilton-extreme phenotype	Artefact by 2/3 checks	Not a robust subtype claim
Five-variable cluster to ADL severity	pseudo-R ² =0.395	Strong apparent association
Four-variable cluster to ADL severity	pseudo-R ² =0.005	Association collapsed after ADL removal

Figures

Audit structure for unsupervised clinical phenotyping



Interpretation: internal fit is not enough; missingness sensitivity and input-outcome circularity must be audited before clinical translation.

Figure 1

Methodological audit structure and input-outcome circularity test. The reference workflow uses routine clinical scale inputs to derive unsupervised clusters, then tests stability under resampling, multiple imputation, and complete-case restriction. The circularity audit removes Barthel ADL from the clustering definition before assessing whether cluster labels still explain ADL severity.

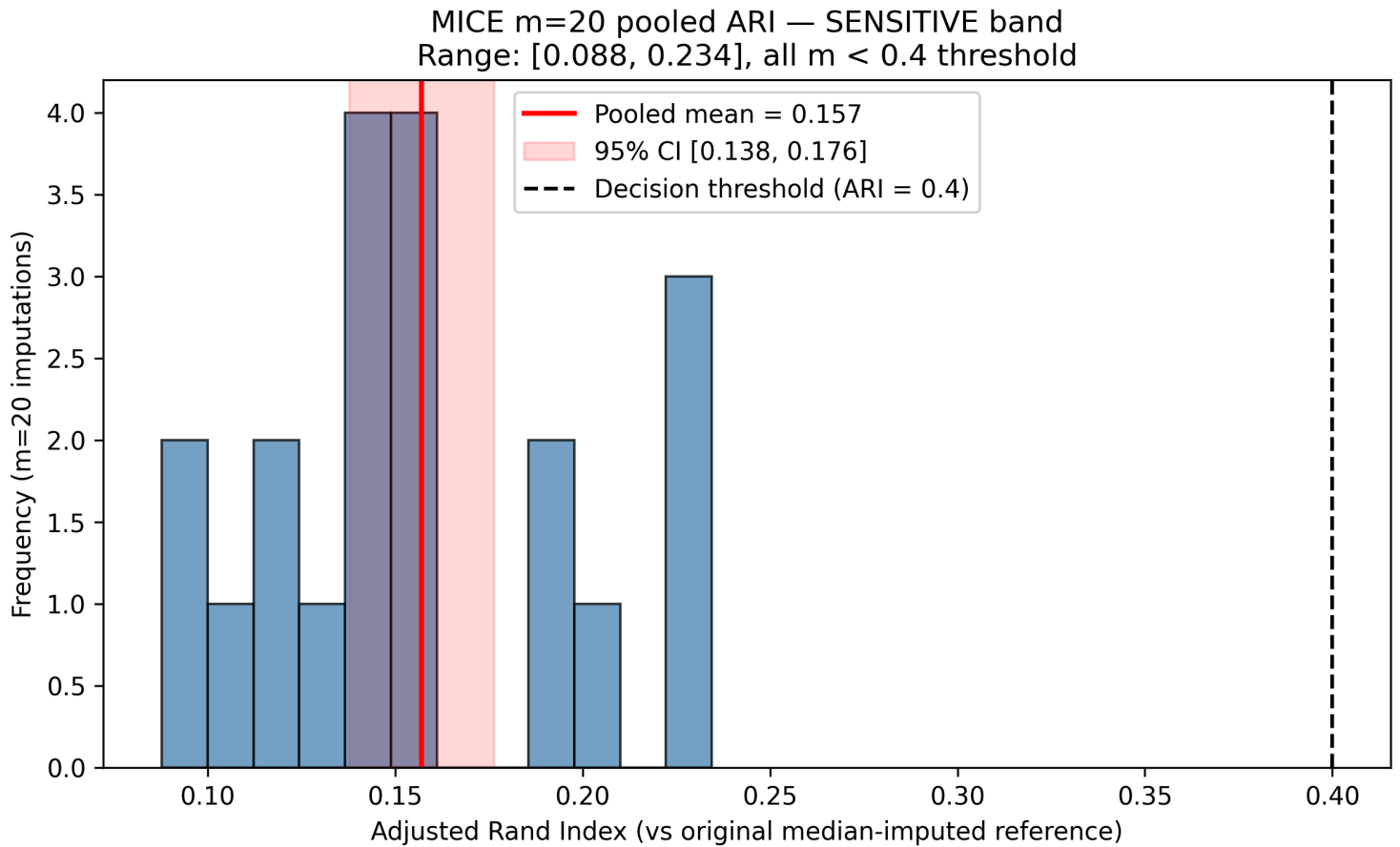


Figure 2

MICE sensitivity of the five-cluster reference partition. Distribution of adjusted Rand index values between each of the 20 MICE-derived five-cluster GMM solutions and the original median-imputed reference partition. The mean ARI was 0.157 (empirical imputation-distribution interval 0.139-0.176), below the pre-specified sensitivity threshold of 0.40.

Supplementary Files

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

- [04supplementarymaterialbmcmrmv2.0.docx](#)