arXiv:2611.02413v1 [cs.LG] 2 Nov 2026

A Scientific Claim Stability Framework for Evaluating MRI Morphometry Pipelines

Duy-Cat Can[1,2,3], Dang-Duat Tran[3], Anh-Khang Nguyen[3], Quoc-Bao Tran[3], Trung-Hieu Do[4,5], Van-Tu Nguyen[6], Duy-Thanh Vu[1,2,3], Oliver Y. Chén[1,2,§], and Hoang-Quynh Le[3,§]

for the Alzheimer's Disease Neuroimaging Initiative*
and the Australian Imaging Biomarkers and Lifestyle flagship study of aging†

[1]Centre for Neural Artificial Intelligence, Lausanne University Hospital, Lausanne, Switzerland
[2]Faculty of Biology and Medicine, University of Lausanne, Lausanne, Switzerland
[3]VNU University of Engineering and Technology, Hanoi, Vietnam
[4]Hanoi Medical University, Hanoi, Vietnam
[5]National Geriatric Hospital, Hanoi, Vietnam
[6]Institute of Diagnostic and Interventional Radiology, Bach Mai Hospital, Hanoi, Vietnam

duy-cat.can@chuv.ch, olivery.chen@chuv.ch
lhquynh@vnu.edu.vn
[§]Corresponding authors

Abstract

MRI morphometry pipelines are often evaluated by measurement agreement or downstream accuracy, yet their scientific value depends on the conclusions they support. Two pipelines can produce similar feature-level results yet support different conclusions about aging, cognitive decline, or neurodegeneration. We introduce Scientific Claim Stability (SCS), a claim-centric evaluation framework that converts region-of-interest (ROI) morphometry tables into auditable claim cards and evaluates their support under changes in dataset, processing pipeline, atlas, feature family, and target. Across 10,370 MRI scans from three datasets and five cohorts, SCS generates 2,783 claim cards and evaluates 89 claim-family pairs. The evaluation distinguishes stable, transferable, and fragile claims. FreeSurfer 8–FastSurfer comparisons show high evidence concordance and low anatomical-support overlap under the primary covariate specification. Acquisition-adjusted analyses show that the magnitude of support disagreement depends on the model specification. Structured review by specialists in geriatrics, neuroscience, and radiology, alongside auxiliary ChatGPT and Gemini reviewers, assesses claim plausibility, clarity, faithfulness, usefulness, and caution. SCS provides a reproducible way to audit the evidential stability of morphometry claims without assuming a universal gold standard.

1 Introduction

Structural MRI morphometry is widely used to support scientific claims about aging, cognitive decline, and neurodegeneration. Large datasets such as ADNI [1], OASIS-1 [2], and AIBL [3] have

*Part of data used in preparation of this article were obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (adni.loni.usc.edu). As such, the investigators within the ADNI contributed to the design and implementation of ADNI and/or provided data but did not participate in the analysis or writing of this report. A complete listing of ADNI investigators can be found at: ADNI Acknowledgment List

†Part of data used in the preparation of this article was obtained from the Australian Imaging Biomarkers and Lifestyle flagship study of aging (AIBL) funded by the Commonwealth Scientific and Industrial Research Organization (CSIRO) which was made available at the ADNI database (adni.loni.usc.edu). The AIBL researchers contributed data but did not participate in analysis or writing of this report. AIBL researchers are listed at data.aibl.org.au/adni/.

Preprint.

made region-of-interest (ROI) morphometry widely available for biomarker discovery and machine learning. Yet the same images can support different conclusions depending on the processing pipeline, feature family, atlas, covariate model, preprocessing choice, and cohort definition. This sensitivity is well documented: FreeSurfer version changes can alter morphometric measurements [4], alternative pipelines encode different assumptions [5–7], and broader analytic flexibility can change scientific inferences even when individual choices are defensible [8, 9].

Existing evaluations assess morphometric measurement agreement, segmentation accuracy, test–retest reliability, and downstream predictive performance [5, 6, 10, 11]. These assessments characterize pipeline outputs, while scientific conclusions also depend on which regional associations support each statement. Two pipelines may have high ROI-level evidence concordance while selecting different brain regions to support a diagnosis, cognition, or age-related statement. The remaining question is whether the same scientific claim remains supported, transferable, and interpretable under realistic dataset, pipeline, atlas, feature-family, and target perturbations.

We introduce Scientific Claim Stability (SCS), a claim-centric framework for evaluating MRI morphometry pipelines. SCS uses scientific claims as its unit of evaluation. It converts ROI morphometry tables into structured claim cards that link a natural-language statement to evidence, support regions, weights, provenance metadata, stability and transfer profiles, and review status. For example, a card describing an association between cortical thickness in a named brain region and MMSE records the supporting statistics and analysis setting. Claim cards extend the structured documentation used in model cards and datasheets [12, 13] to scientific conclusions. We implement `scs_morph`, a configuration-driven Python package for claim extraction, cross-setting evaluation, and expert review.

We evaluate SCS on 10,370 MRI scans from ADNI, OASIS-1, and AIBL. The framework generates 121 claim families and 2,783 claim cards, then evaluates 89 claim-family pairs across dataset transfer, weak-label transfer, pipeline drift, atlas sensitivity, and feature-family sensitivity. The audit identifies a central failure mode: *high evidence concordance does not guarantee scientific claim stability*. In the FreeSurfer–FastSurfer volume comparison under the primary covariate specification, evidence vectors remained highly concordant while supporting regions had very low overlap, making all eight evaluated pairs fragile. Additional acquisition-adjusted analyses show that the magnitude of support disagreement depends on the analysis specification. Claim-level evaluation thus identifies differences in reliability that evidence agreement alone can miss.

This paper makes five contributions. **First**, we formalize scientific claims as typed, transferable evaluation objects extracted from ROI morphometry tables. **Second**, we introduce claim cards and multidimensional stability profiles for documenting interpretable morphometry conclusions. **Third**, we provide `scs_morph`, a reusable framework for claim extraction, evaluation, and expert/LLM review. **Fourth**, we audit 2,783 generated claim cards and 89 evaluated claim-family pairs across datasets, pipelines, atlases, feature families, and targets. **Fifth**, we evaluate selected generated claims through structured human expert and LLM review. Figure 1 summarizes the proposed shift from measurement-level evaluation to claim-level evaluation.

2 Related Work and Evaluation Gap

MRI morphometry pipelines and drift. Structural MRI morphometry pipelines convert anatomical scans into quantitative features such as cortical thickness, cortical regional volume, and subcortical volume. FreeSurfer is widely used for surface-based morphometry [5], while SPM/CAT-style workflows provide alternative voxel-based and morphometric analyses [7, 14]. FastSurfer and SynthSeg further extend the ecosystem with deep-learning-based processing designed for speed or robustness across imaging conditions [6, 15]. However, prior work shows that morphometry estimates can vary across software versions, computing environments, sites, scanners, and quality-control procedures [4, 10, 11]. FreeSurfer version changes can affect regional measures and case-control statistical outcomes [16], and FreeSurfer–CAT12 comparisons show that high within-pipeline reliability does not guarantee strong between-pipeline reproducibility at the ROI level [14, 17]. These studies establish the need for pipeline evaluation through analyses of measurements, reliability, and compatibility; SCS extends their scope to the stability of the claims supported by those measurements.

Analytic flexibility and partially observed multiverses. Neuroimaging results depend on a range of defensible analytic choices [8]. Specification-curve and multiverse analyses make this dependence explicit by examining how conclusions change across reasonable specifications [9, 18]. Recent work

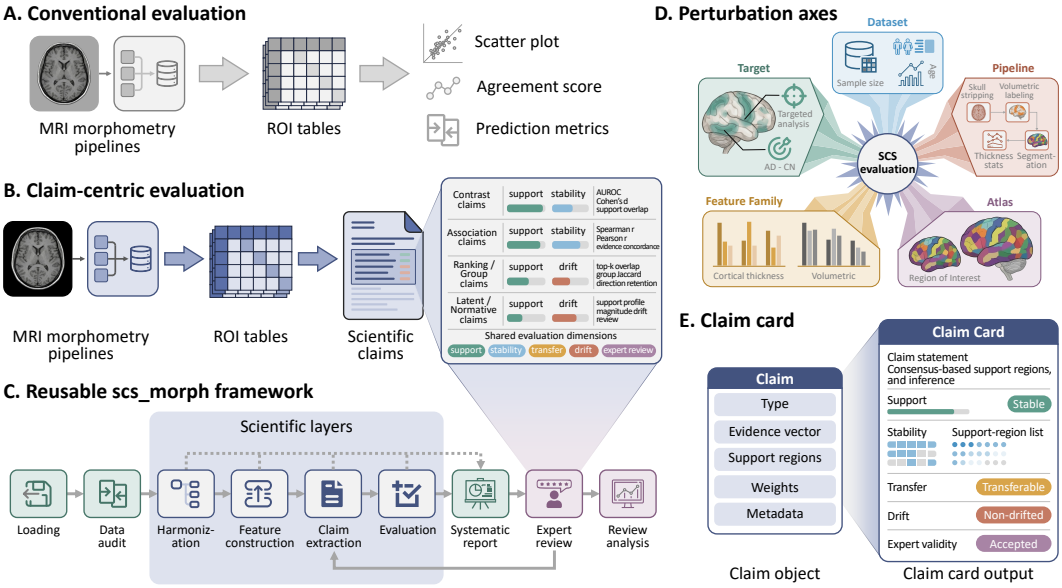


Figure 1: Overview of the Scientific Claim Stability framework. Conventional evaluation compares ROI measurements or prediction scores, whereas SCS converts ROI tables into claim cards and evaluates whether the resulting scientific conclusions remain supported under dataset, pipeline, atlas, feature-family, and target perturbations. Details in Appendix F.1.1.

on multiverse sampling addresses the difficulty of exhaustive enumeration when the space of possible pipelines is large [19]. SCS adopts this perspective but adapts it to a practical morphometry setting in which the multiverse is only partially observed: not every dataset is available under every pipeline, atlas, feature family, target, or covariate specification. SCS identifies stable, transferable, and fragile claims and quantifies their stability under the perturbations that can actually be evaluated.

Documentation artifacts and human-centered evaluation. Model cards and datasheets established structured documentation for model behavior and dataset provenance [12, 13]. Recent work extends this idea to dataset and benchmark documentation, including Croissant for ML-ready dataset metadata and BenchmarkCards for benchmark transparency [20, 21]. SPHERE similarly frames evaluation cards as a way to document human-AI evaluation design and validity [22]. SCS extends this documentation approach to scientific conclusions. Each claim card links a conclusion to evidence, supporting brain regions, perturbation stability, transfer behavior, and expert or LLM review. The scientific claim itself becomes the object of documentation and evaluation.

3 Scientific Claim Stability Framework

3.1 The observed morphometry multiverse

A morphometry conclusion depends on the complete analysis specification, including the dataset. We represent one specification as

$$s \in \mathcal{S} = \mathcal{D} \times \mathcal{H} \times \mathcal{P} \times \mathcal{F} \times \mathcal{A} \times \mathcal{T} \times \mathcal{C}, \tag{1}$$

where $\mathcal{D}$ indexes datasets or views, $\mathcal{H}$ cohorts or subsets, $\mathcal{P}$ pipelines, $\mathcal{F}$ feature families, $\mathcal{A}$ atlases or region systems, $\mathcal{T}$ scientific targets, and $\mathcal{C}$ covariate specifications. These choices jointly define the evaluation object, consistent with the dependence of scientific conclusions on analytic specifications [8, 9].

Each setting s induces an aligned analysis table

$$\Psi(s) = (\mathbf{X}_s, \mathbf{Z}_s, \mathbf{y}_s), \qquad \mathbf{X}_s \in \mathbb{R}^{n_s \times p_s}, \quad \mathbf{Z}_s \in \mathbb{R}^{n_s \times q_s}, \tag{2}$$

where $\mathbf{X}_s$ is the scan-by-region morphometry matrix, $\mathbf{Z}_s$ contains metadata and covariates, and $\mathbf{y}_s$ is an optional target such as diagnosis, MMSE, age, acquisition group, or CDR-derived impairment.

The ideal multiverse $\mathcal{M}$ contains all defensible settings, whereas the observed multiverse

$$\mathcal{M}_{\text{obs}} = \{s_1, \ldots, s_m\} \subset \mathcal{M} \subseteq \mathcal{S} \tag{3}$$

contains only settings that are computationally, institutionally, and empirically available. SCS evaluates stability under the perturbations available in this partially observed multiverse and records unavailable or insufficient comparisons. Operationally, the observed multiverse is a graph: each node is an available analysis setting, and each evaluated edge connects settings through a declared dataset, pipeline, atlas, feature-family, or target change. Here, 121 claim-family settings define 89 predefined evaluation edges, with inference restricted to observed combinations. Appendix B.1 gives the full setting schema, feature-space expansion rules, and matching definitions.

3.2 Claim cards and extraction

SCS defines the scientific claim as the unit of evaluation. For a setting $s \in \mathcal{M}_{\text{obs}}$, a claim card is

$$c_s = (\tau, q, \mathbf{e}, S_k, \mathbf{w}, M), \tag{4}$$

where τ is the claim type, q is the natural-language statement, $\mathbf{e} \in \mathbb{R}^{p_s}$ is a signed evidence vector, $S_k \subseteq \{1, \ldots, p_s\}$ is the support set, $\mathbf{w} \in \mathbb{R}^{p_s}$ is a claim weight vector, and M stores provenance, covariates, feature-space metadata, extraction criteria, and status information. The human view of the card is $\text{view}_{\text{human}}(c_s) = q$. The machine view is $\text{view}_{\text{machine}}(c_s) = (\mathbf{e}, S_k, \mathbf{w}, M)$. The two views link each statement to its evidence, support regions, and evaluation profile.

Claim cards are generated by a typed evidence operator,

$$\mathcal{E}_\tau : (\mathbf{X}_s, \mathbf{y}_s, \mathbf{Z}_s) \mapsto c_s. \tag{5}$$

For supervised claim types, SCS fits ROI-wise covariate-adjusted association models,

$$x_{ij} = \alpha_j + \beta_j y_i + \mathbf{z}_i^\top \boldsymbol{\gamma}_j + \epsilon_{ij}, \qquad i = 1, \ldots, n_s, \qquad j = 1, \ldots, p_s, \tag{6}$$

where ϵ_{ij} denotes the residual error for scan i and region j. The fitted target association is converted into signed regional evidence $e_j = T(\hat{\beta}_j)$, where T may return a signed coefficient, standardized coefficient, effect size, or t-statistic. The support set $S_k(c_s)$ contains regions with the strongest evidence and is checked against explicit support criteria. A simplified support indicator is

$$\mathbb{I}_{\text{support}}(c_s) = \mathbb{I}\left[|S_k(c_s)| \geq k_{\min}, \quad \overline{|e|}_{S_k} \geq \eta, \quad n_s \geq n_{\min}\right], \tag{7}$$

where $\overline{|e|}_{S_k}$ is the mean absolute evidence over the support set. The implementation also supports FDR thresholds, group-level support, direction consistency, and claim-level status assignment. The same claim-card abstraction covers latent-axis and normative-deviation claims, where $\mathbf{e}$ is derived from unsupervised morphometric variation or deviation from a reference group. Appendix B.3 gives the full claim schema and taxonomy, and Appendix C.1 gives the statistical extraction details.

3.3 Stability and transfer profiles

SCS evaluates a source claim c_s against a target setting s_t through a multidimensional profile,

$$\mathcal{V} : (c_s, s_t) \mapsto \boldsymbol{\phi}(c_s, s_t), \qquad \boldsymbol{\phi} = (\phi_{\text{sup}}, \phi_{\text{stab}}, \phi_{\text{trans}}, \phi_{\text{drift}}, \phi_{\text{review}}). \tag{8}$$

SCS reports each component of the profile separately. A claim may be stable but not transferable, transferable but anatomically drifted, or fragile despite high evidence concordance. The metrics are accompanied by descriptive labels such as stable, moderate, fragile, stable-and-transferable, stable-not-transferable, transferable-but-drifted, and insufficient.

For matched feature spaces, SCS separates evidence concordance (EC), support overlap (SO), and direction retention (DR):

$$\text{EC}(s, s_t) = \rho_{\text{Spearman}}\left(\mathbf{e}_s[\mathcal{J}_{st}], \mathbf{e}_t[\mathcal{J}_{st}]\right), \qquad \text{SO}_k(s, s_t) = \frac{|S_k(c_s) \cap S_k(c_t)|}{|S_k(c_s) \cup S_k(c_t)|}, \tag{9}$$

$$\text{DR}_k(s, s_t) = \frac{1}{|S_k(c_s)|} \sum_{j \in S_k(c_s)} \mathbb{I}\left[\text{sign}(e_{s,j}) = \text{sign}(e_{t,j})\right]. \tag{10}$$

Here $\mathcal{J}_{st}$ is the matched ROI set, $\mathbf{e}_s$ and $\mathbf{e}_t$ are the source and target evidence vectors, $e_{s,j}$ and $e_{t,j}$ are their signed evidence values for region j, and c_t is the corresponding target claim. Evidence concordance measures whether the signed evidence ranking is preserved, support overlap measures whether the regions justifying the claim are preserved, and direction retention measures whether source-support regions retain their effect direction. A high global evidence correlation can coexist with a fragile scientific conclusion when the support set or direction pattern changes. Stability permits changes in effect magnitude when the evidence ranking, supporting regions, and effect directions remain sufficiently preserved. The categorical thresholds are preset in the reported experiments and configurable in the software; the continuous metric profile remains the primary result (Appendix C.8).

For non-identical feature spaces, SCS maps source and target regions to a shared group space,

$$g_s : \{1, \ldots, p_s\} \to \mathcal{G}, \qquad g_t : \{1, \ldots, p_t\} \to \mathcal{G}, \tag{11}$$

and computes stability over shared groups such as lobes, network-like groups, ROI groups, or anatomical groups. These mappings allow atlas and feature-family sensitivity to be evaluated in a shared group space.

Transfer is evaluated by applying the source claim as a fixed scoring functional in the target setting. With a normalized source weight vector $\mathbf{w}_s$ restricted to source–target matched features, the transferred claim score for target scan i is

$$u_i^{(t)} = \mathbf{x}_{t,i}^\top \mathbf{w}_s. \tag{12}$$

Binary targets are evaluated with oriented AUROC and effect size (Cohen's d), while continuous targets are evaluated with correlation between $\mathbf{u}^{(t)}$ and $\mathbf{y}_t$. Transfer tests whether a claim extracted in one setting remains useful in another with its weights held fixed. Appendix C gives exact matching, group matching, stability metrics, transfer metrics, and label thresholds.

3.4 Expert and LLM review

SCS also evaluates the scientific usability of generated claims through structured review. For each selected claim c, a human expert or auxiliary LLM reviewer r assigns scores for plausibility, clarity, faithfulness, usefulness, and confidence (higher scores indicate better ratings or greater confidence):

$$a_{r,d}(c) \in \{1, \ldots, 5\}, \qquad d \in \{\text{plausibility}, \text{clarity}, \text{faithfulness}, \text{usefulness}, \text{confidence}\}. \tag{13}$$

Reviewers also assign a caution level and recommended action: accept, revise, reject, or not sure. The review profile summarizes mean scores, reviewer agreement, high-disagreement flags, corrections, and decisions. LLM judges provide scalable auxiliary review, with agreement assessed against human expert scores. The review outcomes are stored with the claim profile alongside statistical support, stability, and transfer measures. Appendix E gives the review prompt, scoring rubric, agreement metrics, and decision rules.

4 Implementation and Empirical Design

4.1 Implementation artifact

We implemented SCS as a configuration-driven Python framework, `scs_morph`. It converts harmonized MRI morphometry tables into claim cards and claim families, evaluates claim-family pairs across observed perturbations, and exports selected cards for expert and LLM review. The workflow includes data audit, metadata harmonization, feature-matrix construction, claim extraction, stability and transfer evaluation, and review analysis. It produces serialized claim cards, claim-family summaries, ROI-level statistics, evaluation-pair results, metric-specific summaries, and review tables. The public implementation is available at `https://github.com/catcd/scs-morph`. The repository contains the code, configurations, and scripts needed to reproduce the workflow after authorized access to the source datasets. Appendix D.5 describes the package structure, configuration files, output artifacts, and script-based reproducibility workflow.

4.2 Datasets, targets, and feature spaces

We evaluate SCS on 10,370 MRI scans from the ADNI (adni.loni.usc.edu) [1], OASIS-1 [2], and AIBL [3] datasets. Established in 2003 as a public–private partnership, ADNI investigates whether

serial MRI, PET, biological markers, and clinical and neuropsychological assessments can measure the progression of MCI and early AD. Dataset-specific access and provenance are documented in Appendix A.2. The empirical multiverse contains full and baseline dataset views, plus an ADNI subset with matched FreeSurfer and SPM/CAT12 cortical-thickness features. ADNI and AIBL are used with CN/MCI/AD-compatible labels, MMSE, age, sex, and acquisition metadata. OASIS-1 uses CDR-derived cognitive impairment, MMSE, age, sex, and education, with a target definition distinct from the harmonized diagnosis labels.

The feature spaces span volumetric and cortical-thickness morphometry. Volume features include FreeSurfer 8 cortical regional volumes and subcortical volumes [5]. Cortical-thickness features include FreeSurfer 8 Schaefer 200/7-network thickness [23], Kong 200/17-network thickness [24], and Yale Brain Atlas thickness [25]. Pipeline perturbation analyses use SPM/CAT12 Schaefer cortical-thickness features [7, 14] and FastSurfer cortical and subcortical volumes [6]. The ADNI SPM-matched view contains 2,892 scans with aligned FreeSurfer 8 and SPM/CAT12 Schaefer cortical-thickness matrices. Appendix D gives the dataset-view inventory, feature-space inventory, matched matrix sizes, and data-access constraints.

4.3 Claim generation and evaluation design

Claims are generated at the analysis-setting level. Each dataset–feature–target specification produces one claim family, and each claim family can contain multiple claim cards at different granularities. Claim extraction produced 121 claim families and 2,783 claim cards. Claim levels include composite, atomic ROI, group-summary, ranking, direction-summary, latent-axis, latent-axis association, normative-deviation, and normative atomic ROI claims. Claim text is generated from structured templates, with each statement linked to source-setting metadata, ROI-level evidence, support regions, and support criteria.

We evaluated 89 claim-family pairs across six perturbation experiments. `E1` evaluates ADNI-to-AIBL volume transfer. `E2` evaluates ADNI-to-OASIS-1 volume transfer, including weak-label transfer from ADNI CN–AD claims to OASIS-1 CDR-derived impairment. `E3` evaluates FreeSurfer 8 versus SPM/CAT12 cortical-thickness drift on the matched ADNI subset. `E4` evaluates FreeSurfer 8 versus FastSurfer volume drift on identical ADNI baseline scans, target definitions, filters, and primary covariates (age, sex, education, and intracranial volume, as applicable). `E5` evaluates FreeSurfer 8 cortical-thickness atlas sensitivity across Schaefer, Kong, and Yale representations. `E6` evaluates FreeSurfer 8 volume-versus-thickness feature-family sensitivity. Exact ROI matching is used when source and target settings share comparable feature names, and group-level matching is used when atlases or feature families do not admit one-to-one ROI correspondence. All 89 configured evaluation pairs were evaluated with no skipped pairs and no processing errors. Appendix D.3 provides the claim-generation inventory, and Appendix D.4 gives the full `E1`–`E6` design and matching rules.

4.4 Expert and LLM review design

The review workflow evaluates whether generated claim cards are plausible, clear, faithful to the displayed evidence, useful for interpretation, and appropriately cautious. From the pool of 2,783 claim cards, the human expert review set contains 60 claims, all of which were also included in the 120-claim LLM review set along with 60 additional claims. The three human reviewers, covering neuroscience, geriatrics, and radiology, each assessed the same 60 claims. LLM judges provide auxiliary review, and their agreement with human experts is evaluated empirically. Reviewers score plausibility, clarity, faithfulness, usefulness, and confidence on 1–5 scales, and also provide a caution level, recommended action, optional corrected claim text, and optional comments. Appendix E gives the full review protocol, prompt, scoring rubric, agreement metrics, and decision rules.

5 Results

5.1 SCS generates an auditable claim landscape

SCS generated 121 claim families and 2,783 claim cards from the observed morphometry multiverse. Figure 2 summarizes the resulting claim landscape across dataset views, feature spaces, claim levels, grouped claim types, and evaluation pairs. The cards cover diagnosis, age, cognition, confounder associations, and unsupervised structure. Atomic ROI claims formed the largest layer with 2,081

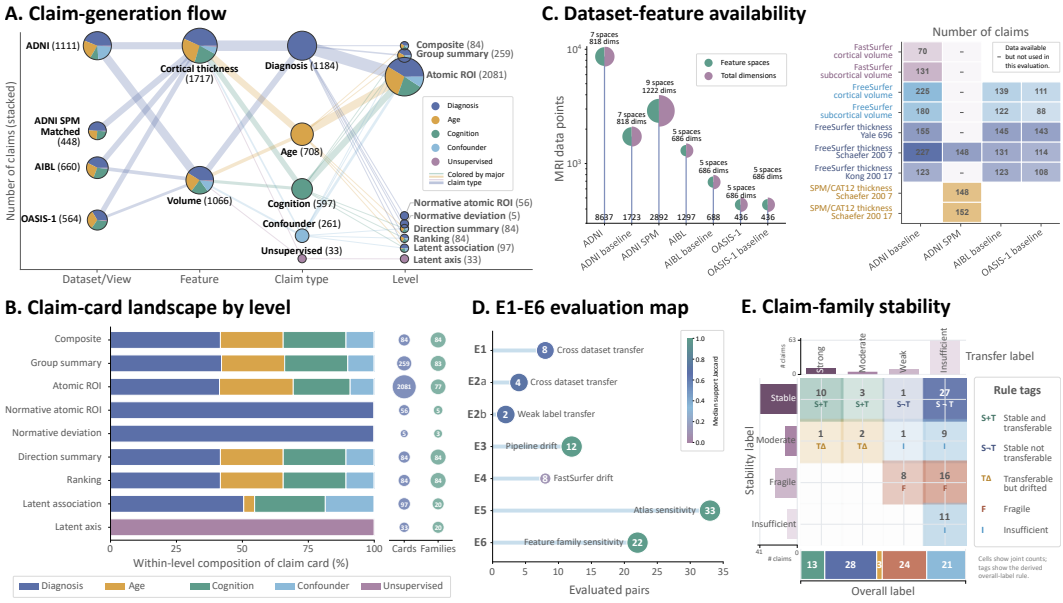


Figure 2: Claim landscape and evaluation coverage. SCS generated 2,783 claim cards from 121 claim families and evaluated 89 claim-family pairs across six experiments. Panels show claim flow, levels, feature availability, experiment coverage, and stability–transfer labels. Details in Appendix F.1.2.

cards, preserving fine-grained regional evidence, while composite, group-summary, ranking, direction-summary, latent-axis, and normative-deviation claims provided higher-level summaries for transfer evaluation and expert review.

Coverage varied by feature family: cortical-thickness features contributed 1,717 claim cards because they span multiple atlases and pipeline comparisons, whereas the 1,066 volume cards primarily support cross-dataset transfer and FreeSurfer–FastSurfer drift analyses. The feature-availability map shows how this coverage relates to the measurements available within each dataset view. Across the landscape, 89 claim-family pairs were evaluated under dataset, pipeline, atlas, and feature-family perturbations. The resulting map records which scientific claims can be generated and which perturbations can be evaluated, linking the claim inventory to the available comparisons.

5.2 Stability and transfer define distinct reliability regimes

The 89 evaluated claim-family pairs differed in both stability and transfer. Across all experiments, 41 pairs were stable, 13 moderate, 24 fragile, and 11 insufficient by stability label. However, stability did not imply transfer: only 13 pairs were stable-and-transferable, whereas 28 were stable-not-transferable and 3 were transferable-but-drifted. Evidence structure, support-region preservation, and claim-score transfer capture different aspects of reliability.

Table 1 summarizes the experiment-level profiles. ADNI-to-AIBL and ADNI-to-OASIS-1 volume transfer accounted for 11 of the 13 stable-and-transferable pairs, the largest contribution among the experiments. The FreeSurfer 8 versus SPM/CAT12 thickness comparison was mostly stable at the evidence and support level, but most pairs were stable-not-transferable because target transfer was unavailable, weak, or insufficient. FreeSurfer 8 versus FastSurfer volume drift had the highest median evidence concordance among the experiments, yet all 8 pairs were fragile, with a median support Jaccard of 0.098. Atlas and feature-family comparisons were more heterogeneous: coarse support was preserved for some targets, while other comparisons were fragile or insufficient.

5.3 Pipeline drift exposes high-concordance fragile-support failures

Figure 3 shows how evidence concordance can coexist with unstable anatomical support. Evidence concordance and support overlap formed distinct reliability axes: stable-and-transferable claims occupied the high-evidence, high-support regime, whereas fragile claims included cases with preserved

Table 1: Experiment-level SCS evaluation summary. EC denotes median evidence concordance (Spearman); SO denotes median support overlap (Jaccard).

Exp.	Perturbation	Pairs	Stability	Overall labels*	EC	SO	Transfer
E1	ADNI→AIBL volume	8	6 stable; 2 moderate	6 S+T; 1 TΔ; 1 I	0.891	0.633	AUROC 0.718; $\lvert\rho\rvert$ 0.352
E2	ADNI→OASIS-1 volume†	6	5 stable; 1 moderate	5 S+T; 1 TΔ	0.842	0.703	AUROC 0.741; $\lvert\rho\rvert$ 0.634
E3	FreeSurfer 8 vs. SPM/CAT12 thickness	12	10 stable; 2 moderate	8 S¬T; 2 S+T; 1 TΔ; 1 I	0.875	1.000	AUROC 0.704; $\lvert\rho\rvert$ 0.363
E4	FreeSurfer 8 vs. FastSurfer volume	8	8 fragile	8 F	0.903	0.098	AUROC 0.561; $\lvert\rho\rvert$ 0.137
E5	FreeSurfer 8 atlas sensitivity‡	33	16 stable; 7 moderate; 10 fragile	16 S¬T; 10 F; 7 I	0.700	1.000	n/a
E6	FreeSurfer 8 volume vs. thickness‡	22	4 stable; 1 moderate; 6 fragile; 11 insufficient	4 S¬T; 6 F; 12 I	0.500	1.000	n/a

Notes. *S+T = stable-and-transferable; S¬T = stable-not-transferable; TΔ = transferable-but-drifted; F = fragile; I = insufficient. †OASIS-1 transfer uses CDR-derived impairment as a weak-label target. ‡SO is computed in grouped anatomical or network space.

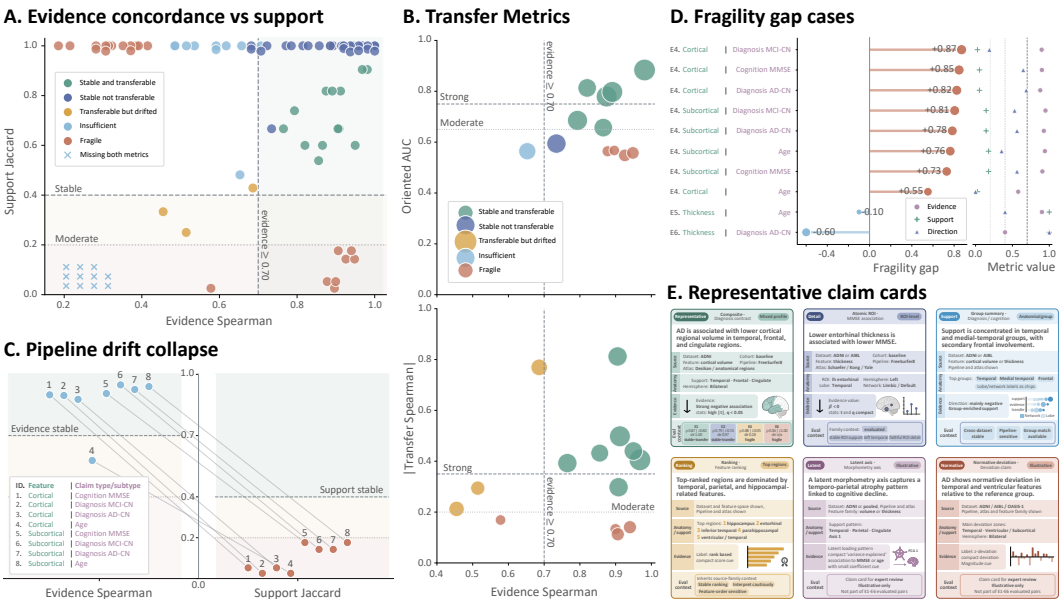


Figure 3: Stability-transfer geometry and fragile support. Evidence concordance and support overlap define distinct claim-reliability regimes. Under the primary E4 covariate specification, all 8 FreeSurfer 8 versus FastSurfer volume pairs were fragile despite high median evidence Spearman, with substantial changes in supporting regions; Appendix F.5 reports sensitivity to support rules and covariate adjustment. Details in Appendix F.1.3.

global evidence ranking but unstable support regions. The clearest example was the FreeSurfer 8 versus FastSurfer volume comparison under the primary E4 specification. All 8 pairs were fragile, with median evidence Spearman of 0.903, support Jaccard of 0.098, and source-support direction retention of 0.546. This pattern covered cortical and subcortical volume and four targets, although cortical age showed lower evidence concordance (0.578). An exact hypergeometric lower-tail test of matched-size support overlap, BH-adjusted over the eight comparisons, found below-random overlap for all four cortical pairs ($q = 0.012$–0.040); the four subcortical pairs were compatible with random overlap ($q = 0.448$–0.563).

Anatomical-support agreement varied with sample selection and covariate adjustment (Table 2). Under the primary covariate specification, restricting the analysis to acquisition-complete observations increased median support Jaccard from 0.098 to 0.194. Further adjustment for field strength and cohort on these same observations raised it to 0.510, with direction retention reaching 1.000; adding site fixed effects yielded a Jaccard of 0.404, although the CN–AD site models were rank deficient. E4 support agreement therefore depends on the analysis specification, even when evidence concordance is high. Additional support-rule, overlap-null, bootstrap, and covariate analyses appear in Appendix F.5.

Table 2: E4 sensitivity to sample selection and covariate adjustment (medians across eight FreeSurfer 8–FastSurfer claim-family comparisons). EC: evidence concordance (Spearman); SO: support overlap (Jaccard); DR: direction retention over source-support regions.

Analysis specification	**EC**	**SO**	**DR**
Primary covariates, full sample	0.903	0.098	0.546
Primary covariates, acquisition-complete subset	0.885	0.194	0.656
Additional adjustment for field strength and cohort	0.892	0.510	1.000
Additional adjustment for site fixed effects	0.724	0.404	1.000

Note. Primary covariates are age, sex, education, and intracranial volume, as applicable. The final three rows use identical acquisition-complete observations within each pair, with covariates added cumulatively. CN–AD site models are rank deficient.

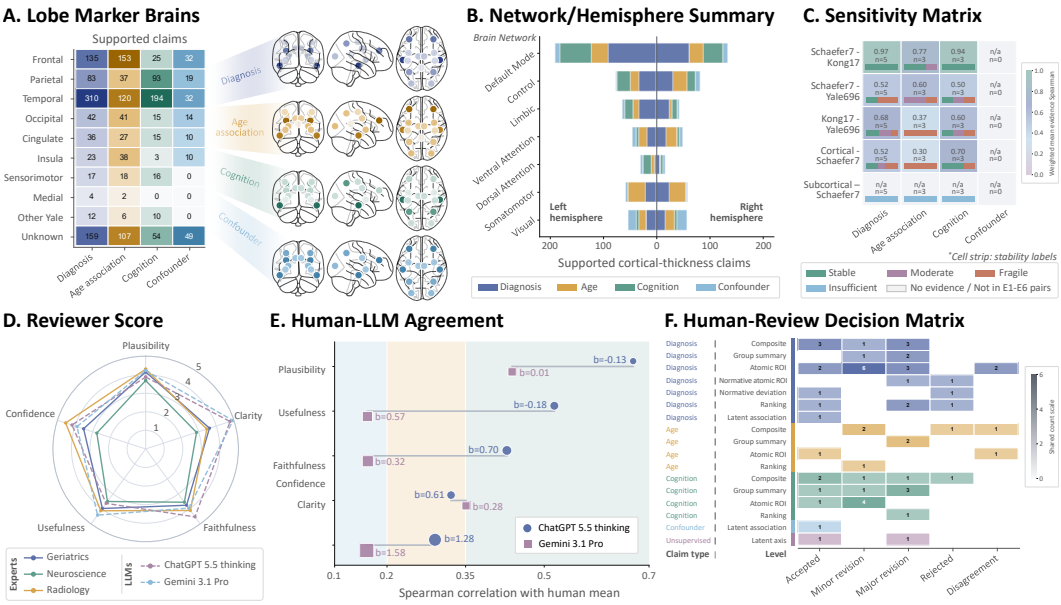


Figure 4: Anatomical, representation, and review summaries. Panels A–C summarize supported claims across anatomical groups and representation changes. Panels D–F summarize human/LLM review scores, agreement, and decisions. The anatomical panels summarize the distribution of supporting regions across generated claims. Details in Appendix F.1.4.

5.4 Representation sensitivity, anatomy, and review

Figure 4 summarizes anatomical support, representation sensitivity, and expert/LLM review. Supported claims frequently involved anatomical groups consistent with established neurodegeneration findings. Temporal regions were the most prominent lobe-level support group across diagnosis/impairment and cognition targets, and Default-network cortical-thickness claims were the most frequent network-level group. These patterns are consistent with prior reports of AD-related cortical thinning in temporal, parietal, precuneus, cingulate, and frontal regions, and with default-mode network vulnerability in aging and AD [26–28]. The hemisphere summary is exploratory because AD-related asymmetry varies across anatomical regions and studies [29, 30].

The representation-sensitivity panel compares claim stability across atlases and feature families. Cortical-thickness atlas comparisons retained more stable or moderate structure than volume-versus-thickness comparisons, whereas subcortical-volume-to-thickness rows were frequently insufficient because the feature spaces lacked a meaningful shared cortical ROI mapping. Diagnosis/impairment, cognition, and age targets also showed different sensitivity patterns. Anatomical plausibility and stability under atlas or feature-family changes are evaluated separately.

Structured review assessed the scientific usability of selected generated claim cards. Among the 60 human-reviewed claims, 15 were accepted, 17 required minor revision, 19 required major revision, 5 were rejected, and 4 showed expert disagreement. Mean plausibility and faithfulness were approxi-

mately 4.05 on the 1–5 scale, while mean clarity, usefulness, and confidence were approximately 4.20, 3.95, and 3.85. Human reviewer agreement was moderate and dimension-dependent, with ICC$(2, k)$ values of 0.690 for plausibility, 0.665 for clarity, and 0.628 for faithfulness, but lower agreement for usefulness and confidence. Among the LLM judges, ChatGPT had the strongest agreement with human plausibility scores (Spearman correlation 0.670), while both LLM reviewers tended to score clarity higher than the human mean. Expert comments identified opaque ROI/atlas terminology, insufficient caution, overbroad anatomical summaries, mixed-direction aggregation, and misclassified global or QC features; some rejected claims misidentified features or the direction of an association. The review identified corrections needed for scientific use of statistically generated candidate claims.

6 Discussion and Limitations

SCS evaluates MRI morphometry pipelines through the scientific claims researchers report. It extends conventional evaluations of measurement agreement, segmentation accuracy, and downstream prediction by examining whether scientific conclusions remain supported, stable, transferable, and usable under realistic perturbations. Each claim is represented as a typed, machine-evaluable object with a natural-language statement, signed evidence vector, support regions, transfer behavior, perturbation profile, and review status. Across 10,370 MRI scans, SCS generated 2,783 claim cards and evaluated 89 claim-family pairs across dataset transfer, weak-label transfer, pipeline drift, atlas sensitivity, and feature-family sensitivity. These experiments show that claim-level auditing is feasible across the evaluated settings and perturbations without a universal gold standard.

The central empirical finding is the high-concordance fragile-support failure mode under the primary FreeSurfer 8 versus FastSurfer covariate specification. The four cortical support overlaps were lower than a matched-size random-overlap baseline after BH correction; the four subcortical overlaps were statistically compatible with chance. On the acquisition-complete subset, median support Jaccard increased from 0.194 to 0.510 after adding field strength and cohort to the covariate model for the same observations. The full analysis specification, including the dataset, pipeline, covariates, and inclusion criteria, is therefore recorded as the scientific claim's provenance. SCS records evidence concordance, support overlap, effect direction, and transfer, together with configurable descriptive labels. These measures assess preservation of evidence and anatomical support under a specified perturbation; biological truth and clinical validity require separate evidence.

The observed multiverse comprises 121 analyzed claim-family settings and 89 predefined comparisons, with incomplete coverage of possible combinations. In neuroimaging studies, datasets differ in the available pipelines, atlases, feature families, targets, and covariate specifications. SCS evaluates claim stability over the available comparisons and records unavailable, weakly matched, or insufficient comparisons. This is relevant for studies that reuse processed data, integrate cohorts with different access constraints, compare tools across subsets, or evaluate related but non-identical targets such as diagnosis-derived claims and CDR-derived impairment.

The `scs_morph` implementation separates data harmonization, feature construction, claim extraction, stability evaluation, transfer testing, and expert/LLM review into reusable modules. This modular design allows new datasets, atlases, pipelines, targets, covariate models, and review protocols to be added without changing the core claim-card abstraction. Human review assesses claim cards for plausibility, clarity, faithfulness, usefulness, and caution. LLM judges provide scalable auxiliary scores, with agreement assessed against human experts.

The current study has several limitations. **First**, this paper focuses on ROI-level morphometry tables; vertex-wise cortical thickness, full surface geometry, voxel-wise maps, curvature, sulcal depth, and mesh-level quality are outside its empirical scope. **Second**, the empirical analysis is cross-sectional; longitudinal processing, visit-level dependence, and claim stability of atrophy rates are left for future work. **Third**, ADNI and AIBL participant-level source and derived data remain subject to their access and redistribution restrictions; we provide code and aggregate results for independent regeneration under authorized access. OASIS-1 has its own data-use terms and is also omitted from our participant-level artifact release. **Fourth**, OASIS-1 uses CDR-derived weak labels, whereas ADNI and AIBL use harmonized CN/MCI/AD labels. OASIS transfer therefore evaluates a related clinical target through weak-label transfer. **Fifth**, the expert-review study reports agreement and revision decisions for a targeted set of claims across three clinical and scientific specialties, limiting conclusions about clinical consensus. Claim-level evaluation tests whether scientific support is preserved under a declared change in dataset, pipeline, representation, target, or covariate specification.

Acknowledgments and Disclosure of Funding

Funding. This work was supported by the Vietnam National Foundation for Science and Technology Development (NAFOSTED) under the project "Research and Development of a Personalized Machine Learning System for Early Alzheimer's Disease Diagnosis from Adaptive Multimodal Biomarkers" (Grant No. 102.05-2025.54).

Duy-Thanh Vu and Oliver Y. Chén gratefully acknowledge support from the Swiss National Science Foundation (SNSF; Project No. 3200–0–239967 and Project No. CR00I5–235987).

ADNI data acknowledgment. Data collection and sharing for the Alzheimer's Disease Neuroimaging Initiative (ADNI) is funded by the National Institute on Aging (National Institutes of Health Grant U19AG024904). The grantee organization is the Northern California Institute for Research and Education. In the past, ADNI has also received funding from the National Institute of Biomedical Imaging and Bioengineering, the Canadian Institutes of Health Research, and private sector contributions through the Foundation for the National Institutes of Health (FNIH) including generous contributions from the following: AbbVie, Alzheimer's Association; Alzheimer's Drug Discovery Foundation; Araclon Biotech; BioClinica, Inc.; Biogen; Bristol-Myers Squibb Company; CereSpir, Inc.; Cogstate; Eisai Inc.; Elan Pharmaceuticals, Inc.; Eli Lilly and Company; EuroImmun; F. Hoffmann-La Roche Ltd and its affiliated company Genentech, Inc.; Fujirebio; GE Healthcare; IXICO Ltd.; Janssen Alzheimer Immunotherapy Research & Development, LLC.; Johnson & Johnson Pharmaceutical Research & Development LLC.; Lumosity; Lundbeck; Merck & Co., Inc.; Meso Scale Diagnostics, LLC.; NeuroRx Research; Neurotrack Technologies; Novartis Pharmaceuticals Corporation; Pfizer Inc.; Piramal Imaging; Servier; Takeda Pharmaceutical Company; and Transition Therapeutics.

AIBL data acknowledgment. Data used in the preparation of this article were obtained from the Australian Imaging Biomarkers and Lifestyle flagship study of ageing (AIBL). See `https://www.aibl.csiro.au` for further details. Data were collected by the AIBL study group. AIBL study methodology has been reported previously [3].

OASIS-1 data acknowledgment. Data were provided in part by OASIS-1: Cross-Sectional. OASIS-1 principal investigators include D. Marcus, R. Buckner, J. Csernansky, and J. Morris. OASIS-1 was supported by P50 AG05681, P01 AG03991, P01 AG026276, R01 AG021910, P20 MH071616, and U24 RR021382.

Competing interests. The authors declare no competing interests.

References

[1] Ronald C. Petersen, Paul S. Aisen, Laurel A. Beckett, Michael C. Donohue, Anthony C. Gamst, Danielle J. Harvey, Clifford R. Jack, William J. Jagust, Leslie M. Shaw, Arthur W. Toga, John Q. Trojanowski, and Michael W. Weiner. Alzheimer's Disease Neuroimaging Initiative (ADNI): Clinical Characterization. *Neurology*, 74(3):201–209, January 2010.

[2] Daniel S. Marcus, Tracy H. Wang, Jamie Parker, John G. Csernansky, John C. Morris, and Randy L. Buckner. Open Access Series of Imaging Studies (OASIS): Cross-Sectional MRI Data in Young, Middle Aged, Nondemented, and Demented Older Adults. *Journal of Cognitive Neuroscience*, 19(9):1498–1507, September 2007.

[3] Kathryn A. Ellis, Ashley I. Bush, David Darby, Daniela De Fazio, Jonathan Foster, Peter Hudson, Nicola T. Lautenschlager, Nat Lenzo, Ralph N. Martins, Paul Maruff, Colin Masters, Andrew Milner, Kerryn Pike, Christopher Rowe, Greg Savage, Cassandra Szoeke, Kevin Taddei, Victor Villemagne, Michael Woodward, and David Ames. The Australian Imaging, Biomarkers and Lifestyle (AIBL) Study of Aging: Methodology and Baseline Characteristics of 1112 Individuals Recruited for a Longitudinal Study of Alzheimer's Disease. *International Psychogeriatrics*, 21(4):672–687, August 2009.

[4] Ed H. B. M. Gronenschild, Petra Habets, Heidi I. L. Jacobs, Ron Mengelers, Nico Rozendaal, Jim van Os, and Machteld Marcelis. The Effects of FreeSurfer Version, Workstation Type, and Macintosh Operating System Version on Anatomical Volume and Cortical Thickness Measurements. *PLOS ONE*, 7(6):e38234, June 2012.

[5] Bruce Fischl. FreeSurfer. *NeuroImage*, 62(2):774–781, August 2012.

[6] Leonie Henschel, Sailesh Conjeti, Santiago Estrada, Kersten Diers, Bruce Fischl, and Martin Reuter. FastSurfer: A Fast and Accurate Deep Learning Based Neuroimaging Pipeline. *NeuroImage*, 219:117012, October 2020.

[7] John Ashburner and Karl J. Friston. Voxel-Based Morphometry—The Methods. *NeuroImage*, 11(6):805–821, June 2000.

[8] Joshua Carp. On the Plurality of (Methodological) Worlds: Estimating the Analytic Flexibility of fMRI Experiments. *Frontiers in Neuroscience*, 6:149, October 2012.

[9] Uri Simonsohn, Joseph P. Simmons, and Leif D. Nelson. Specification Curve Analysis. *Nature Human Behaviour*, 4:1208–1214, 2020.

[10] Zafer Iscan, Tony B. Jin, Alexandria Kendrick, Bryan Szeglin, Hanzhang Lu, Madhukar Trivedi, Maurizio Fava, Patrick J. McGrath, Myrna Weissman, Benji T. Kurian, Phillip Adams, Sarah Weyandt, Marisa Toups, Thomas Carmody, Melvin McInnis, Cristina Cusin, Crystal Cooper, Maria A. Oquendo, Ramin V. Parsey, and Christine DeLorenzo. Test–Retest Reliability of FreeSurfer Measurements Within and Between Sites: Effects of Visual Approval Process. *Human Brain Mapping*, 36(9):3472–3485, September 2015.

[11] Graham N. Knussmann, Jeffrey S. Anderson, Molly B. D. Prigge, Douglas C. Dean, Nicholas Lange, Erin D. Bigler, Andrew L. Alexander, Janet E. Lainhart, Brandon A. Zielinski, and Jace B. King. Test–Retest Reliability of FreeSurfer-Derived Volume, Area and Cortical Thickness from MPRAGE and MP2RAGE Brain MRI Images. *NeuroImage: Reports*, 2(2):100086, June 2022.

[12] Margaret Mitchell, Simone Wu, Andrew Zaldivar, Parker Barnes, Lucy Vasserman, Ben Hutchinson, Elena Spitzer, Inioluwa Deborah Raji, and Timnit Gebru. Model Cards for Model Reporting. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, pages 220–229. Association for Computing Machinery, January 2019.

[13] Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé, III, and Kate Crawford. Datasheets for Datasets. *Communications of the ACM*, 64(12):86–92, November 2021.

[14] Maria de Fátima Machado Dias, Paulo Carvalho, Miguel Castelo-Branco, and João Valente Duarte. Cortical Thickness in Brain Imaging Studies Using FreeSurfer and CAT12: A Matter of Reproducibility. *NeuroImage: Reports*, 2(4):100137, December 2022.

[15] Benjamin Billot, Douglas N. Greve, Oula Puonti, Axel Thielscher, Koen Van Leemput, Bruce Fischl, Adrian V. Dalca, and Juan Eugenio Iglesias. SynthSeg: Segmentation of Brain MRI Scans of Any Contrast and Resolution Without Retraining. *Medical Image Analysis*, 86:102789, May 2023.

[16] Pavel Filip, Petr Bednarik, Lynn E. Eberly, Amir Moheet, Alena Svatkova, Heidi I. Gröhn, A. F. Kumar, Elizabeth R. Seaquist, and Silvia Mangia. Different FreeSurfer Versions Might Generate Different Statistical Outcomes in Case–Control Comparison Studies. *Neuroradiology*, 64(4):765–773, April 2022.

[17] Benita Schmitz-Koep, Vivian Schultz, Fabian Bongratz, Aurore Menegaux, Melissa Thalhammer, Severin Schramm, Su Hwan Kim, Claus Zimmer, Christian Sorg, Christian Wachinger, Panteleimon Giannakopoulos, Marie-Louise Montandon, Cristelle Rodriguez, Sven Haller, and Dennis M. Hedderich. Longitudinal Assessment of Cortical Thickness in Healthy Older Individuals: A Comparison Between CAT12 and FreeSurfer. *Neuroradiology*, 68:469–482, February 2026. Published online 12 December 2025.

[18] Sara Steegen, Francis Tuerlinckx, Andrew Gelman, and Wolf Vanpaemel. Increasing transparency through a multiverse analysis. *Perspectives on Psychological Science*, 11(5):702–712, September 2016.

[19] Cassie Ann Short, Andrea Hildebrandt, Robin Bosse, Stefan Debener, Metin Özyağcılar, Katharina Paul, Jan Wacker, and Daniel Kristanto. Lost in a Large EEG Multiverse? Comparing Sampling Approaches for Representative Pipeline Selection. *Journal of Neuroscience Methods*, 424:110564, December 2025.

[20] Mubashara Akhtar, Omar Benjelloun, Costanza Conforti, Luca Foschini, Pieter Gijsbers, Joan Giner-Miguelez, Sujata Goswami, Nitisha Jain, Michalis Karamousadakis, Satyapriya Krishna, Michael Kuchnik, Sylvain Lesage, Quentin Lhoest, Pierre Marcenac, Manil Maskey, Peter Mattson, Luis Oala, Hamidah Oderinwale, Pierre Ruyssen, Tim Santos, Rajat Shinde, Elena Simperl, Arjun Suresh, Goeff Thomas, Slava Tykhonov, Joaquin Vanschoren, Susheel Varma, Jos van der Velde, Steffen Vogler, Carole-Jean Wu, and Luyao Zhang. Croissant: A Metadata Format for ML-Ready Datasets. In *Advances in Neural Information Processing Systems*, volume 37, 2024.

[21] Anna Sokol, Elizabeth M. Daly, Michael Hind, David Piorkowski, Xiangliang Zhang, Nuno Moniz, and Nitesh V. Chawla. BenchmarkCards: Standardized Documentation for Large Language Model Benchmarks. In *Advances in Neural Information Processing Systems*, 2025. NeurIPS 2025 Datasets and Benchmarks Track.

[22] Dora Zhao, Qianou Ma, Xinran Zhao, Chenglei Si, Chenyang Yang, Ryan Louie, Ehud Reiter, Diyi Yang, and Tongshuang Wu. SPHERE: An Evaluation Card for Human-AI Systems. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar, editors, *Findings of the Association for Computational Linguistics: ACL 2025*, pages 1340–1365, Vienna, Austria, July 2025. Association for Computational Linguistics.

[23] Alexander Schaefer, Ru Kong, Evan M. Gordon, Timothy O. Laumann, Xi-Nian Zuo, Avram J. Holmes, Simon B. Eickhoff, and B. T. Thomas Yeo. Local–Global Parcellation of the Human Cerebral Cortex from Intrinsic Functional Connectivity MRI. *Cerebral Cortex*, 28(9):3095–3114, September 2018.

[24] Ru Kong, Qing Yang, Evan M. Gordon, Aihuiping Xue, Xiaoyang Yan, Csaba Orban, Xi-Nian Zuo, R. Nathan Spreng, Tian Ge, Avram J. Holmes, Simon B. Eickhoff, and B. T. Thomas Yeo. Individual-Specific Areal-Level Parcellations Improve Functional Connectivity Prediction of Behavior. *Cerebral Cortex*, 31(10):4477–4500, October 2021.

[25] Evan Collins, Omar Chishti, Hari McGrath, Sami Obaid, Alex King, Edwin Qiu, Ellie Gabriel, Xilin Shen, Jagriti Arora, Xenophon Papademetris, R. Todd Constable, Dennis D. Spencer, and Hitten P. Zaveri. Yale Brain Atlas to Interactively Explore Multimodal Structural and Functional Neuroimaging Data. *Frontiers in Network Physiology*, 5:1585019, June 2025.

[26] Bradford C. Dickerson, Akram Bakkour, David H. Salat, Eric Feczko, Jenni Pacheco, Douglas N. Greve, Francine Grodstein, Christopher I. Wright, Deborah Blacker, Herminia D. Rosas, Reisa A. Sperling, Alireza Atri, John H. Growdon, Bradley T. Hyman, John C. Morris, Bruce Fischl, and Randy L. Buckner. The Cortical Signature of Alzheimer's Disease: Regionally Specific Cortical Thinning Relates to Symptom Severity in Very Mild to Mild AD Dementia and Is Detectable in Asymptomatic Amyloid-Positive Individuals. *Cerebral Cortex*, 19(3):497–510, March 2009.

[27] Katell Mevel, Gaël Chételat, Francis Eustache, and Béatrice Desgranges. The Default Mode Network in Healthy Aging and Alzheimer's Disease. *International Journal of Alzheimer's Disease*, 2011:535816, June 2011.

[28] Sara L. Seoane, Martijn van den Heuvel, Ángel Acebes, and Niels Janssen. The Subcortical Default Mode Network and Alzheimer's Disease: A Systematic Review and Meta-Analysis. *Brain Communications*, 6(2):fcae128, April 2024.

[29] James M. Roe, Didac Vidal-Piñeiro, Øystein Sørensen, Andreas M. Brandmaier, Sandra Düzel, Hector A. Gonzalez, Rogier A. Kievit, Ethan Knights, Simone Kühn, Ulman Lindenberger, Athanasia M. Mowinckel, Lars Nyberg, Denise C. Park, Sara Pudas, Melissa M. Rundle, Kristine B. Walhovd, Anders M. Fjell, and René Westerhausen. Asymmetric Thinning of the Cerebral Cortex Across the Adult Lifespan Is Accelerated in Alzheimer's Disease. *Nature Communications*, 12(1):721, February 2021.

[30] Lora Minkova, Annegret Habich, Jessica Peter, Christoph P. Kaller, Simon B. Eickhoff, and Stefan Klöppel. Gray Matter Asymmetries in Aging and Neurodegeneration: A Review and Meta-Analysis. *Human Brain Mapping*, 38(12):5890–5904, December 2017.

[31] Rahul S. Desikan, Florent Ségonne, Bruce Fischl, Brian T. Quinn, Bradford C. Dickerson, Deborah Blacker, Randy L. Buckner, Anders M. Dale, R. Paul Maguire, Bradley T. Hyman, Marilyn S. Albert, and Ronald J. Killiany. An Automated Labeling System for Subdividing the Human Cerebral Cortex on MRI Scans into Gyral Based Regions of Interest. *NeuroImage*, 31(3):968–980, July 2006.

[32] Bruce Fischl, André van der Kouwe, Christophe Destrieux, Eric Halgren, Florent Ségonne, David H. Salat, Evelina Busa, Larry J. Seidman, Jill Goldstein, David Kennedy, Verne Caviness, Nikos Makris, Bruce Rosen, and Anders M. Dale. Automatically Parcellating the Human Cerebral Cortex. *Cerebral Cortex*, 14(1):11–22, January 2004.

[33] Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer Series in Statistics. Springer, New York, NY, 2 edition, 2009.

[34] Yoav Benjamini and Yosef Hochberg. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society: Series B (Methodological)*, 57(1):289–300, January 1995.

[35] Ian T. Jolliffe. *Principal Component Analysis*. Springer Series in Statistics. Springer, New York, NY, 2 edition, 2002.

[36] Andre F. Marquand, Iead Rezek, Jan Buitelaar, and Christian F. Beckmann. Understanding Heterogeneity in Clinical Cohorts Using Normative Models: Beyond Case–Control Studies. *Biological Psychiatry*, 80(7):552–561, October 2016.

[37] Patrick E. Shrout and Joseph L. Fleiss. Intraclass Correlations: Uses in Assessing Rater Reliability. *Psychological Bulletin*, 86(2):420–428, March 1979.

[38] Kenneth O. McGraw and S. P. Wong. Forming Inferences About Some Intraclass Correlation Coefficients. *Psychological Methods*, 1(1):30–46, March 1996.

A Code, Data, and Artifact Availability

A.1 Code availability

The SCS framework is released as a reproducible code repository:

`https://github.com/catcd/scs-morph`

The repository contains the `scs_morph` package, configuration files, command-line scripts, documentation, and aggregate evaluation summaries needed to reproduce the SCS workflow after obtaining authorized access to the required source datasets. ADNI and AIBL participant-level data and derived participant-level artifacts are excluded from the release, as are OASIS-1 participant-level data and features. The dataset-specific conditions are described below. Researchers must obtain the source datasets through the corresponding data-access procedures and regenerate restricted participant-level artifacts locally under the applicable data-use agreements.

A.2 Data availability

This study uses ADNI, AIBL, and OASIS-1 data. The ADNI and AIBL source data used in this evaluation were downloaded from the LONI Image and Data Archive (IDA) on January 14, 2026; the respective archives comprise separately updated imaging and metadata files. The released repository omits participant-level MRI data, clinical variables, scan identifiers, raw feature matrices, derived participant-level morphometry tables, and participant-level claim scores. Authorized researchers can regenerate the corresponding analyses after obtaining each source dataset under its applicable terms.

ADNI data are available to approved researchers through the LONI Image and Data Archive after application and acceptance of the ADNI Data Use Agreement.[3] The ADNI Data Use Agreement prohibits redistribution of individual participant-level raw or derived data and permits publication of summary, non-participant-level results. It also restricts use of AI tools on ADNI participant-level data unless the tools provide guarantees that the data are safeguarded, contained, and not released to other parties. Accordingly, this work does not release ADNI participant-level data and does not provide ADNI participant-level data, scan-level features, identifiers, or clinical rows to public-facing AI systems.

AIBL data access is governed by the AIBL Data Use Agreement through the LONI/ADNI access infrastructure.[4] AIBL data are used only for scientific research purposes, are not redistributed, and are not included as participant-level or derived participant-level records in the released repository.

OASIS-1 data are available under the OASIS data-use terms.[5] These terms restrict use to non-commercial academic research without prior authorization, prohibit re-identification attempts and facial-recognition or reconstruction uses, and require dataset acknowledgment. This study releases aggregate analyses and does not include OASIS-1 participant-level MRI data or derived feature matrices in its public artifacts.

A.3 Released artifacts

The released repository includes the framework and configuration files required to reproduce the SCS workflow. The released artifacts include:

- configuration files defining datasets, feature spaces, claim extraction, evaluation pairs, thresholds, and expert-review settings;
- scripts for data audit, harmonization, feature-matrix construction, claim extraction, claim evaluation, and expert-review analysis;
- documentation describing expected input locations and output artifacts;
- aggregate result summaries, evaluation-pair summaries, and claim-level summaries that contain no participant identifiers, no scan-level rows, no individual clinical records, and no participant-level feature vectors.

[3] `https://adni.loni.usc.edu/data-samples/adni-data/`

[4] `https://ida.loni.usc.edu/collaboration/access/appApply.jsp?project=AIBL`

[5] `https://sites.wustl.edu/oasisbrains/home/access/`

A.4 Reproducibility scope

Authorized users can reproduce the workflow by obtaining the source datasets through the corresponding access procedures, placing the required raw metadata and morphometry files under the configured raw-data directory, and running the scripts described in Appendix D.5. Participant-level artifacts, including clinical tables, and scan-level features, remain subject to the corresponding data-use agreements. The repository supports reproduction of the SCS methodology and aggregate evaluation outputs under dataset-specific restrictions on participant-level and derived participant-level data.

B Framework Details and Claim Objects

B.1 Analysis settings and observed multiverse

The main text defines an analysis setting as an element of the product space $\mathcal{S} = \mathcal{D} \times \mathcal{H} \times \mathcal{P} \times \mathcal{F} \times \mathcal{A} \times \mathcal{T} \times \mathcal{C}$. The implementation uses the following expanded definition. An analysis setting is a reproducible specification

$$s = (d, h, p, f, a, t, c, r, \kappa), \tag{14}$$

where d is the dataset view, h is the cohort or subset, p is the processing pipeline, f is the feature family, a is the atlas or region system, t is the target, c is the covariate specification, r is the inclusion rule, and κ denotes claim-generation and support criteria. The main text uses the shorter notation $s = (d, h, p, f, a, t, c)$; the implementation also retains r and κ to specify each claim family fully. Table B1 summarizes the analysis-setting fields and their corresponding implementation names.

Table B1: Analysis-setting fields used by SCS. Each setting defines one reproducible morphometry specification from which claim families can be extracted.

Field	Symbol	Implementation	Meaning
dataset or view	D	`dataset`	Dataset or derived view, including ADNI [1], AIBL [3], OASIS-1 [2], baseline views, and matched views such as the ADNI FreeSurfer–SPM/CAT12 matched subset.
cohort or subset	H	`cohort`, view tag	Cohort, data split, or subset identifier, such as ADNI1, ADNI-GO/2, ADNI3, baseline-only views, pooled views, matched subsets, or cohort-specific analysis groups.
pipeline	P	`pipeline`	Processing pipeline or software family used to derive morphometric features, including FreeSurfer 8 [5], Statistical Parametric Mapping (SPM) with CAT12 [7, 14], and FastSurfer [6].
feature family	F	`feature_family`	Type of morphometric measurement, including subcortical volume, cortical regional volume, and cortical thickness.
atlas or region system	A	`atlas`	Region system used to organize ROI features, including Schaefer [23], Kong [24], Yale Brain Atlas [25], and FreeSurfer anatomical regions [31, 32].
target	T	`target`	Scientific target or contrast used to extract claims, including diagnosis, MMSE, age, CDR-derived impairment, demographic variables, or acquisition variables.
covariates	C	`covariates`	Adjustment variables included in the ROI-wise association model, such as age, sex, education, intracranial volume, or other setting-specific covariates.
inclusion rule	R	`inclusion_filter`	Row-level eligibility rule used to define the analysis population before fitting claim-extraction models.
criteria	κ	`criteria`, thresholds	Claim-generation and support rules, including minimum sample size, top-k support size, effect-strength thresholds, FDR thresholds, and status-label thresholds.

Each setting induces an aligned analysis table

$$\Psi(s) = (\mathbf{X}_s, \mathbf{Z}_s, \mathbf{y}_s), \tag{15}$$

where $\mathbf{X}_s \in \mathbb{R}^{n_s \times p_s}$ is the morphometry matrix, $\mathbf{Z}_s \in \mathbb{R}^{n_s \times q_s}$ contains metadata and covariates, and $\mathbf{y}_s \in \mathcal{Y}_s^{n_s}$ is the target vector when a supervised target is available. Rows correspond to scans or scan-level records, and columns of $\mathbf{X}_s$ correspond to anatomical regions or region-level features. The alignment operator removes rows with missing target, missing covariates, or non-finite feature values when required by the setting. For unsupervised or normative settings, $\mathbf{y}_s$ may be absent or used only for downstream interpretation.

The observed multiverse is the set of settings instantiated by available data, feature spaces, and experiment configurations:

$$\mathcal{M}_{\text{obs}} = \{s_i : i = 1, \ldots, m\}. \tag{16}$$

This set is a subset of the ideal multiverse $\mathcal{M}$ of all defensible analysis specifications. In practice, $\mathcal{M}_{\text{obs}}$ is partially observed because not every dataset is processed by every pipeline, not every atlas exists for every feature family, and not every target is available in every cohort. SCS evaluates claim stability over observed perturbations and marks unsupported comparisons.

B.2 Feature spaces and setting expansion

A feature space specifies how a column in the harmonized data is converted into a matrix $\mathbf{X}_s$. Formally, a feature space is represented as

$$\omega = (\mathrm{name}, \mathrm{column}, f, p, a, \mathcal{R}_\omega), \tag{17}$$

where name is the feature-space identifier, column is the source column containing the vectorized features, f is the feature family, p is the pipeline, a is the atlas or region system, and $\mathcal{R}_\omega$ is the ordered set of feature names. Separating the feature matrix from its provenance allows the same extraction procedure to operate on cortical volumes, subcortical volumes, and cortical-thickness atlases.

Setting expansion creates analysis settings by combining dataset views, feature spaces, targets, covariates, and claim-generation rules. Let Ω denote the set of feature spaces and $\mathcal{Q}$ the set of claim queries. A configuration expansion can be written as

$$\mathrm{Expand}(\mathcal{D}_{\mathrm{obs}}, \Omega, \mathcal{Q}) \to \mathcal{M}_{\mathrm{obs}}. \tag{18}$$

Each generated setting must satisfy availability and sample-size constraints. For a setting s, the eligibility indicator is

$$\mathbb{I}_{\mathrm{eligible}}(s) = \mathbb{I}\left[n_s \geq n_{\min}\right] \mathbb{I}\left[p_s \geq p_{\min}\right] \mathbb{I}\left[\mathbf{y}_s \text{ is available when required}\right], \tag{19}$$

where $n_{\min}$ is the minimum sample size and $p_{\min}$ is the minimum number of usable features. Settings that fail these constraints are skipped or marked as insufficient.

Evaluation pairs are ordered pairs of settings,

$$\pi = (s_{\mathrm{src}}, s_{\mathrm{tgt}}, \delta, \mu), \tag{20}$$

where s_{src} is the source setting, s_{tgt} is the target setting, δ is the perturbation relation, and μ is the feature-matching strategy. The perturbation relation records what changes between settings, such as dataset, weak-label target, pipeline, atlas, feature family, or target. The matching strategy μ determines whether evidence vectors are compared through exact ROI identity or through a coarser anatomical grouping.

For exact ROI matching, the shared index set is

$$\mathcal{J}_{st} = \mathcal{R}_{s_{\mathrm{src}}} \cap \mathcal{R}_{s_{\mathrm{tgt}}}, \tag{21}$$

where $\mathcal{R}_s$ is the feature-name set for setting s. For non-identical feature spaces, SCS uses a grouping map

$$g_s : \mathcal{R}_s \to \mathcal{G}, \tag{22}$$

where $\mathcal{G}$ may correspond to anatomical groups, lobes, coarse networks, or network-like groups. The shared comparison space is then

$$\mathcal{G}_{st} = g_{s_{\mathrm{src}}}(\mathcal{R}_{s_{\mathrm{src}}}) \cap g_{s_{\mathrm{tgt}}}(\mathcal{R}_{s_{\mathrm{tgt}}}). \tag{23}$$

This permits atlas sensitivity and feature-family sensitivity analyses even when source and target ROI tables do not share identical feature columns. Exact metric definitions for matched and grouped comparisons are given in Appendix C.6.

B.3 Claim-card schema

A claim card is the serialized artifact corresponding to the formal claim object in Eq. (4). The released implementation defines this artifact in the `ClaimCard` schema and stores claim cards as JSON-serializable records with fields for identity, source setting, anatomy, statistical evidence, support, generated text, and status. Table B2 groups the claim-card fields by function.

Claim cards are validated before serialization. The released implementation checks generated text for empty statements, unresolved placeholders, invalid values such as `nan` or `None`, and unsafe or overclaiming language. If a supported claim fails text validation, the validation routine records the warning and can downgrade the claim to weak support. Text validation checks the statements presented for human review and scientific reporting.

Table B2: Claim-card fields. A claim card stores both the human-readable statement and the machine-readable evidence needed to audit, transfer, and review the claim.

Field group	Fields	Purpose
identity	`claim_id`, `claim_family_id`, `claim_level`, `claim_type`	Uniquely identifies the claim and links it to the family of claims extracted from the same analysis setting.
source setting	`source_setting`, `dataset`, `cohort`, `n_features`, `feature_space`, `feature_family`, `pipeline`, `atlas`, `target`, `covariates`, `n_samples`	Records the dataset view, feature provenance, target, covariate model, and sample/feature dimensions that produced the claim.
anatomy	`roi`, `hemisphere`, `roi_group`, `anatomical_group`, `network`, `lobe`	Places the claim in an anatomical, network, or group-level coordinate system.
evidence	`direction`, `effect_size`, `evidence_metric`, `evidence_value`, `coefficient`, `standard_error`, `t_value`, `p_value`, `q_value`	Stores the signed statistical evidence supporting the claim.
support	`support_regions`, `support_region_effects`, `support_metrics`, `criteria_passed`, `criteria_details`	Records the regions and criteria used to justify the claim.
text	`claim_text`, `short_claim_text`, `machine_readable_statement`, `expert_review_hint`	Links the natural-language claim to a structured statement suitable for evaluation and review.
status	`status`, `notes`	Stores support status and warnings from validation or serialization.

B.4 Claim levels and claim types

SCS separates the *level* of a claim from the *type* of scientific target it addresses. The claim level determines the granularity of the statement, whereas the claim type determines the statistical or scientific question. A diagnosis contrast, MMSE association, age association, latent axis, or normative-deviation analysis can produce composite, regional, group-level, ranking, or direction-summary claims. The supported levels, types, and statuses are defined in the `ClaimCard` schema and used consistently by the claim-family extraction modules. Table B3 lists the supported claim levels, claim types, and claim statuses.

Table B3: Claim levels, claim types, and claim statuses. Claim levels specify the granularity of the statement, claim types specify the scientific target, and statuses specify the support or processing state.

Category	Value	Meaning
claim level	`composite`	A multi-region statement summarizing the main support pattern for one claim family.
claim level	`atomic_roi`	A single-region statement linked to one ROI-level effect.
claim level	`group_summary`	A statement summarizing support concentration in an anatomical group, network, lobe, or related region group.
claim level	`ranking`	A statement describing the ranked support regions for a target or contrast.
claim level	`direction_summary`	A statement describing the dominant sign of effects in the support set.
claim level	`latent_axis`	An unsupervised morphometry-axis statement derived without a supervised target.
claim level	`latent_axis_association`	A statement linking an unsupervised morphometry axis to metadata or labels.
claim level	`normative_deviation`	A statement describing deviation of a target group from a reference group.
claim level	`normative_atomic_roi`	A single-region normative-deviation statement.
claim type	`diagnosis_contrast`	Binary diagnostic contrast such as CN versus AD or CN versus MCI.
claim type	`cdr_impairment_contrast`	CDR-derived cognitive-impairment contrast, for weak-label settings such as OASIS-1.
claim type	`mmse_association`	Association between morphometry and MMSE.
claim type	`age_association`	Association between morphometry and age.
claim type	`continuous_association`	Generic association with a continuous target.
claim type	`demographic_contrast`	Contrast defined by a demographic variable.
claim type	`acquisition_contrast`	Contrast defined by an acquisition or imaging-protocol variable.
claim type	`feature_ranking`	Ranking of regions by evidence strength.
claim type	`group_enrichment`	Enrichment or concentration of support in an anatomical group.
claim type	`direction_summary`	Summary of the dominant direction of supported effects.
claim type	`latent_morphometry_axis`	Unsupervised morphometry axis extracted from the feature matrix.
claim type	`latent_axis_association`	Association between a latent axis and metadata or labels.
claim type	`normative_deviation`	Deviation from a reference distribution or normative group.
claim type	`normative_group_deviation`	Group-level concentration of normative deviations.
status	`generated`	Claim card was created but has not yet been assigned a support status.
status	`supported`	Claim passes the predefined support criteria.
status	`weak_support`	Claim is generated but should be treated as exploratory or downgraded by validation.
status	`criteria_failed`	Claim does not pass support criteria.
status	`skipped`	Claim generation or evaluation was skipped for a documented reason.
status	`error`	Claim generation or evaluation encountered an error.

B.5 Claim templates

Claim text is generated from structured templates with explicit evidence-to-text mappings. Templates are populated with the claim type, target labels, direction, feature-family label, region summary, group label, and covariate text. Each rendered statement is linked to the statistical evidence and source setting. In the released implementation, the template library is specified in `configs/claim_templates.yaml`, while claim-generation settings and support criteria are configured through `configs/claim_generation.yaml` and `configs/claim_extraction_all_main.yaml`. The templates are rendered by the claim text-generation module and validated before serialization. Table B4 gives representative templates used by the framework.

Table B4: Representative claim templates. Templates generate human-readable statements while preserving a direct link to evidence vectors, support regions, and source-setting metadata.

Template class	Claim type or level	Template pattern
diagnosis contrast	`diagnosis_contrast`	`{positive_label} is associated with {direction} {feature_family_label} in {region_summary} compared with {negative_label}, after adjustment for {covariate_text}.`
MMSE association	`mmse_association`	`Higher MMSE is associated with {direction} {feature_family_label} in {region_summary}, after adjustment for {covariate_text}.`
age association	`age_association`	`Older age is associated with {direction} {feature_family_label} in {region_summary}, after adjustment for {covariate_text}.`
ranking	`feature_ranking`	`The top-ranked regions for {target_label} are {top_regions}.`
group enrichment	`group_enrichment`	`The claim signal is concentrated in {group_label}, where {n_support_rois}/{n_group_rois} regions meet support criteria.`
direction summary	`direction_summary`	`Across the support set, {fraction_pct}% of regions show lower/higher {feature_family_label} for {target_label}.`
latent axis	`latent_morphometry_axis`	`An unsupervised morphometry axis explains {explained_variance_pct}% of variance and is dominated by {top_regions}.`
normative deviation	`normative_deviation`	`{target_group} shows normative deviation in {feature_family_label}, most prominently in {region_summary}, relative to the {reference_group} reference.`

Templates constrain the wording of causal and diagnostic statements to the available evidence. Generated text is rejected or flagged when it contains unresolved placeholders, invalid values, unknown anatomical labels, or banned language. The weak-support template is used when a generated claim does not meet the predefined evidence threshold and should be treated as exploratory.

C Claim Extraction and Evaluation Metrics

C.1 ROI-wise association models

For supervised claim families, SCS fits one covariate-adjusted association model per region to estimate regional target associations [33]. Before model fitting, the feature matrix and metadata table are aligned by index or row order, rows with missing target or required covariates are removed, and rows with non-finite feature values are excluded. Categorical covariates are one-hot encoded with a dropped reference level, and numeric covariates are converted to floating-point values. Depending on the setting, features and targets can be standardized before model fitting.

For region j, the regression model is

$$x_{ij} = \alpha_j + \beta_j y_i + \mathbf{z}_i^\top \boldsymbol{\gamma}_j + \epsilon_{ij}, \qquad i = 1, \ldots, n_s, \qquad j = 1, \ldots, p_s. \tag{24}$$

Here x_{ij} is the morphometry value for scan i and region j, y_i is the target variable, and $\mathbf{z}_i$ is the covariate vector. For binary contrasts, $y_i \in \{0, 1\}$ encodes the negative and positive groups. For

continuous associations, y_i is a continuous target such as MMSE or age. The coefficient $\hat{\beta}_j$ is the region-level target association after covariate adjustment.

Let $\mathbf{d}_i = [1, y_i, \mathbf{z}_i^\top]^\top$ be the design vector and let $\mathbf{D} \in \mathbb{R}^{n_s \times r_s}$ be the design matrix. For a full-column-rank design $\mathbf{D}$, ordinary least squares estimates

$$\hat{\boldsymbol{\theta}}_j = (\mathbf{D}^\top \mathbf{D})^{-1} \mathbf{D}^\top \mathbf{x}_j, \qquad \hat{\boldsymbol{\theta}}_j = (\hat{\alpha}_j, \hat{\beta}_j, \hat{\boldsymbol{\gamma}}_j^\top)^\top. \tag{25}$$

The implementation uses the target coefficient and its standard error to compute a t-statistic, p-value, and signed evidence value for each ROI. In the primary extraction procedure, regions that cannot be fitted are excluded from support assessment. Separate categorical site-effect sensitivity models use a pseudoinverse estimator for rank-deficient designs; their coefficients are interpreted only at the level of sensitivity summaries (Appendix F.5).

Table C1 summarizes the extraction models used by the implemented claim families.

Table C1: Claim extraction models. Most claim families are derived from ROI-wise association models, while latent-axis and normative-deviation claims use alternative evidence operators.

Claim type	**Model**	**Target**	**Evidence**	**Support**
diagnosis contrast	ROI-wise OLS with binary target and covariates	CN vs AD, CN vs MCI, or related binary contrast	signed coefficient, t-statistic, p-value, q-value	top-k regions and FDR-supported ROIs
CDR impairment contrast	ROI-wise OLS with binary weak-label target	CDR-derived impairment	signed coefficient, t-statistic, p-value, q-value	top-k regions and FDR-supported ROIs
MMSE association	ROI-wise OLS with continuous target and covariates	MMSE	signed coefficient, t-statistic, p-value, q-value	top-k regions and FDR-supported ROIs
age association	ROI-wise OLS with continuous target and covariates	age	signed coefficient, t-statistic, p-value, q-value	top-k regions and FDR-supported ROIs
demographic or acquisition contrast	ROI-wise OLS with binary or encoded target	sex, field strength, or acquisition group	signed coefficient, t-statistic, p-value, q-value	top-k regions and FDR-supported ROIs
latent morphometry axis	residualization followed by PCA/SVD	none required	component loading and explained variance	top-loading regions
latent-axis association	association between latent score and metadata	post-hoc metadata target	correlation or regression coefficient	metadata-associated latent axis
normative deviation	reference distribution and z-score deviation	target group versus reference group	regional deviation score	top-deviation regions

C.2 Evidence metrics and FDR adjustment

For each fitted ROI-wise model, SCS stores the estimated target coefficient $\hat{\beta}_j$, its standard error $\widehat{\mathrm{se}}(\hat{\beta}_j)$, a t-statistic, and a two-sided p-value. The default signed evidence value is the t-statistic,

$$e_j = t_j = \frac{\hat{\beta}_j}{\widehat{\mathrm{se}}(\hat{\beta}_j)}. \tag{26}$$

Other signed evidence definitions, such as standardized coefficients or signed effect sizes, can be used without changing the claim-card structure. The evidence value e_j must preserve magnitude and direction for consistent evaluation of support, direction, and transfer.

For a family of p_s ROI-wise tests, SCS applies Benjamini–Hochberg false-discovery-rate correction [34]. Let $p_{(1)} \leq \cdots \leq p_{(p_s)}$ be the ordered p-values. The adjusted value for rank r is computed as

$$q_{(r)} = \min_{\ell \geq r} \left\{ \frac{p_s}{\ell} p_{(\ell)} \right\}, \tag{27}$$

with monotonicity enforced over the ordered tests. The resulting q_j values are stored in each ROI statistic and in atomic claim cards when applicable. FDR-adjusted evidence is used for support criteria, but top-k support is retained even when no ROI passes a strict FDR threshold, allowing the framework to distinguish weak exploratory claims from unsupported or failed claims.

C.3 Support selection and criteria

Given an evidence vector $\mathbf{e}$, SCS first ranks regions by absolute evidence. The top-k support set is

$$S_k = \{j : |e_j| \text{ is among the } k \text{ largest values in } \{|e_1|, \ldots, |e_{p_s}|\}\}. \tag{28}$$

For text generation and transfer evaluation, the implementation stores both the support-region names and their signed effects. The support set provides interpretability, while the full evidence vector remains available for stability and transfer metrics.

SCS then applies claim-level criteria. For a generic claim card, the support decision is

$$\mathrm{Pass}(c_s) = \mathbb{I}\left[n_s \geq n_{\min}\right] \mathbb{I}\left[|S_k| \geq k_{\min}\right] \mathbb{I}\left[\overline{|e|}_{S_k} \geq \eta\right] \mathbb{I}\left[N_{\mathrm{FDR}} \geq m_{\min} \text{ or } \mathrm{ExploratoryAllowed}\right], \tag{29}$$

where N_{FDR} is the number of ROIs passing the configured FDR threshold. The exact values of $n_{\min}$, $k_{\min}$, η, and $m_{\min}$ are configuration-dependent. Evidence extraction and support assessment are separate steps: a claim can be generated, weakly supported, supported, or rejected by criteria.

Different claim levels use different criteria, summarized in Table C2. Composite claims require sufficient overall support. Atomic ROI claims require ROI-level statistical support. Group-summary claims require support concentration within an anatomical or network group. Ranking claims require a stable ranked support set. Direction-summary claims require a sufficient fraction of support regions to share the same sign. Latent-axis and normative-deviation claims use criteria based on variance explained, loading magnitude, reference-group size, or deviation thresholds.

Table C2: Support criteria by claim level. Criteria determine whether a generated statement is marked as supported, weakly supported, or criteria-failed.

Claim level	**Minimum criteria**	**Output**
composite	Minimum sample size, non-empty top-k support, sufficient mean absolute evidence, and optional FDR support	Multi-region claim summarizing the main evidence pattern
atomic ROI	ROI-level evidence passes the configured statistical or effect-size threshold	Single-region claim with coefficient, t-statistic, p-value, and q-value
group summary	A group contains enough support ROIs or sufficient group-level mean absolute evidence	Anatomical, lobe, or network-level enrichment claim
ranking	Ranked region list can be generated and meets minimum evidence criteria	Ordered list of regions supporting the target or contrast
direction summary	A configured fraction of support regions shares the same sign	Directional summary of whether support is predominantly positive, negative, or mixed
latent axis	Component explains sufficient variance and has interpretable high-loading regions	Unsupervised morphometry-axis claim
latent-axis association	Latent scores show sufficient association with metadata or labels	Claim linking latent morphometry variation to a target or covariate
normative deviation	Reference group is large enough and target group shows sufficient z-score deviation	Group-level deviation claim relative to a reference distribution
normative atomic ROI	Individual ROI deviation exceeds configured z-threshold	Single-region normative-deviation claim

C.4 Latent-axis and normative-deviation claims

SCS can extract claims without a supervised target and can characterize deviations from a reference group. For latent-axis claims, the feature matrix is first optionally residualized with respect to covariates and standardized across regions. A low-rank decomposition is then applied to the residualized matrix, following the standard principal-component view of dominant multivariate variation [35]. For the first component,

$$\tilde{\mathbf{X}}_s \approx \mathbf{u}_1 \sigma_1 \mathbf{v}_1^\top, \tag{30}$$

where $\mathbf{v}_1 \in \mathbb{R}^{p_s}$ is the loading vector over regions and $\sigma_1^2 / \sum_\ell \sigma_\ell^2$ gives the explained-variance fraction. The loading vector $\mathbf{v}_1$ is used as the evidence vector $\mathbf{e}$, with support defined by the top-loading regions. If metadata or labels are available, the component score $\mathbf{u}_1 \sigma_1$ can be associated post hoc with age, diagnosis, MMSE, or acquisition variables to generate latent-axis association claims.

Normative-deviation claims compare individuals or groups with a reference distribution, following normative modeling approaches in neuroimaging [36]. SCS defines a reference group R and a target group G within a setting. For region j, the reference mean and standard deviation are

$$\mu_{Rj} = \frac{1}{|R|} \sum_{i \in R} x_{ij}, \qquad \sigma_{Rj} = \sqrt{\frac{1}{|R|-1} \sum_{i \in R} (x_{ij} - \mu_{Rj})^2}. \tag{31}$$

The regional deviation of the target group is then

$$z_{Gj} = \frac{\bar{x}_{Gj} - \mu_{Rj}}{\sigma_{Rj} + \epsilon}, \qquad \bar{x}_{Gj} = \frac{1}{|G|} \sum_{i \in G} x_{ij}, \tag{32}$$

where ϵ prevents division by zero. The vector $\mathbf{z}_G = (z_{G1}, \ldots, z_{Gp_s})$ becomes the evidence vector for normative-deviation claims. Support is selected from regions with the largest absolute deviations or from regions exceeding a configured z-threshold. This procedure generates interpretable claims in weakly labeled settings where a full diagnostic target is unavailable.

C.5 Exact and group-level feature matching

Exact ROI matching. For source setting s and target setting t, exact ROI matching is used when both settings share a comparable region system. Let $\mathcal{R}_s$ and $\mathcal{R}_t$ denote the ordered sets of ROI names in the source and target ROI-statistic tables. The matched ROI set is

$$\mathcal{J}_{st} = \mathcal{R}_s \cap \mathcal{R}_t. \tag{33}$$

All exact-match stability metrics are computed only on $\mathcal{J}_{st}$. If $|\mathcal{J}_{st}| < 5$, the comparison is marked as insufficient because evidence concordance and support overlap are not meaningful with too few matched features. Exact matching is used for comparisons such as matched FreeSurfer–SPM Schaefer features or FreeSurfer–FastSurfer features with shared anatomical names.

For a matched ROI $j \in \mathcal{J}_{st}$, the source and target statistics are denoted $e_{s,j}$ and $e_{t,j}$. The default signed evidence is the ROI-wise t-statistic, but the same definitions apply to coefficients or other signed evidence values. Support sets are derived from the source and target ROI-statistic tables using the same top-k rule unless otherwise specified.

Group-level matching for non-identical feature spaces. Some comparisons do not admit exact ROI matching. For example, Schaefer, Kong, and Yale cortical-thickness atlases have different parcels, and regional volume and cortical thickness use different region systems. SCS maps ROI names to a common anatomical or network-level space for group-level matching. Let

$$g_s : \mathcal{R}_s \to \mathcal{G}, \qquad g_t : \mathcal{R}_t \to \mathcal{G} \tag{34}$$

map source and target ROIs to shared groups such as `network_7_like`, `lobe`, `roi_group`, or `anatomical_group`. The shared group set is

$$\mathcal{G}_{st} = g_s(\mathcal{R}_s) \cap g_t(\mathcal{R}_t). \tag{35}$$

For each group $g \in \mathcal{G}_{st}$, SCS aggregates signed evidence by averaging over ROIs assigned to the group:

$$\bar{e}_{s,g} = \frac{1}{|\mathcal{R}_{s,g}|} \sum_{j \in \mathcal{R}_{s,g}} e_{s,j}, \qquad \mathcal{R}_{s,g} = \{j \in \mathcal{R}_s : g_s(j) = g\}. \tag{36}$$

The same operation is applied to the target setting. Group-level support sets are then defined by ranking groups according to $|\bar{e}_{s,g}|$ or group-level mean absolute evidence. The resulting group-level evidence permits comparison across non-identical feature spaces.

C.6 Evidence, support, direction, and magnitude stability

SCS computes stability as a collection of complementary metrics. Table C3 summarizes the metrics used for exact ROI and group-level comparisons. Group-level comparisons use the same metrics, replacing ROI evidence e_j with group evidence $\bar{e}_g$.

Here $R_k(s)$ denotes the top-k ranked ROI set by absolute evidence, and $S_k(c_s)$ denotes the support set attached to the source claim card. The notation $z(\cdot)$ denotes standardization within the compared evidence vector before computing magnitude drift. For group-level metrics, $S_k^{\mathcal{G}}(c_s)$ denotes the top-k support groups.

Table C3: Stability metrics used by SCS. The metrics separate evidence agreement, support-region preservation, direction consistency, and effect-size drift.

Metric	Formula	Interpretation
evidence Spearman	$\rho_{\mathrm{Spearman}}(\mathbf{e}_s[\mathcal{J}_{st}], \mathbf{e}_t[\mathcal{J}_{st}])$	Rank concordance of signed evidence vectors over matched ROIs.
evidence Pearson	$\rho_{\mathrm{Pearson}}(\mathbf{e}_s[\mathcal{J}_{st}], \mathbf{e}_t[\mathcal{J}_{st}])$	Linear concordance of signed evidence vectors over matched ROIs.
weighted Jaccard of absolute evidence	$\frac{\sum_{j\in\mathcal{J}_{st}} \min(\lvert e_{s,j}\rvert, \lvert e_{t,j}\rvert)}{\sum_{j\in\mathcal{J}_{st}} \max(\lvert e_{s,j}\rvert, \lvert e_{t,j}\rvert)}$	Overlap of evidence magnitude profiles, giving larger weight to high-evidence ROIs.
top-k rank overlap	$\frac{\lvert R_k(s) \cap R_k(t)\rvert}{k}$	Fraction of top-ranked evidence regions shared by source and target.
support Jaccard	$\frac{\lvert S_k(c_s) \cap S_k(c_t)\rvert}{\lvert S_k(c_s) \cup S_k(c_t)\rvert}$	Overlap of the source and target support regions.
source support overlap	$\frac{\lvert S_k(c_s) \cap S_k(c_t)\rvert}{\lvert S_k(c_s)\rvert}$	Fraction of source-support regions retained in the target support set.
target support overlap	$\frac{\lvert S_k(c_s) \cap S_k(c_t)\rvert}{\lvert S_k(c_t)\rvert}$	Fraction of target-support regions already present in the source support set.
direction retention	$\frac{1}{\lvert S_k(c_s)\rvert} \sum_{j\in S_k(c_s)} \mathbb{I}[\mathrm{sign}(e_{s,j}) = \mathrm{sign}(e_{t,j})]$	Fraction of source-support regions with preserved effect direction.
sign-flip fraction	$1 - DR_k(s,t)$	Fraction of source-support regions whose effect direction changes sign.
magnitude drift over matched features	$\frac{1}{\lvert\mathcal{J}_{st}\rvert} \sum_{j\in\mathcal{J}_{st}} \lvert z(e_{s,j}) - z(e_{t,j})\rvert$	Mean standardized evidence change over all matched ROIs.
magnitude drift over source support	$\frac{1}{\lvert S_k(c_s)\rvert} \sum_{j\in S_k(c_s)} \lvert z(e_{s,j}) - z(e_{t,j})\rvert$	Mean standardized evidence change over the regions used to justify the source claim.
group evidence concordance	$\rho(\bar{\mathbf{e}}_s[\mathcal{G}_{st}], \bar{\mathbf{e}}_t[\mathcal{G}_{st}])$	Evidence concordance after aggregation to a shared group space.
group support Jaccard	$\frac{\lvert S_k^{\mathcal{G}}(c_s) \cap S_k^{\mathcal{G}}(c_t)\rvert}{\lvert S_k^{\mathcal{G}}(c_s) \cup S_k^{\mathcal{G}}(c_t)\rvert}$	Overlap of high-evidence anatomical, network, or lobe groups.
group direction retention	$\frac{1}{\lvert S_k^{\mathcal{G}}(c_s)\rvert} \sum_{g\in S_k^{\mathcal{G}}(c_s)} \mathbb{I}[\mathrm{sign}(\bar{e}_{s,g}) = \mathrm{sign}(\bar{e}_{t,g})]$	Direction preservation over source-support groups.

C.7 Claim-score transfer

Claim-score transfer evaluates whether a source claim remains useful in target data without refitting. The source claim defines a weight vector $\mathbf{w}_s$ over matched target features. By default, the weight vector is derived from source coefficients or signed evidence values and normalized:

$$\mathbf{w}_s = \frac{\tilde{\mathbf{e}}_s}{\|\tilde{\mathbf{e}}_s\|_2 + \epsilon}, \tag{37}$$

where $\tilde{\mathbf{e}}_s$ is the source evidence restricted to matched features. For target feature matrix $\mathbf{X}_t$, the transferred claim score is

$$\mathbf{u}^{(t)} = \mathbf{X}_t \mathbf{w}_s. \tag{38}$$

The target feature matrix is standardized before scoring when configured.

Table C4 summarizes the transfer metrics. For binary targets, SCS orients the AUROC so that values above 0.5 indicate alignment with the configured positive class. Cohen's d is computed as the standardized mean difference between claim scores in the positive and negative target groups. For continuous targets, SCS reports both Spearman and Pearson correlations between the transferred claim score and the target.

Table C4: Transfer metrics used by SCS. Transfer evaluates whether a source claim remains useful in target data without re-estimating the claim.

Target type	Metric	Interpretation
binary	oriented AUROC	Discrimination of the target contrast by the transferred claim score, oriented so higher values favor the configured positive class.
binary	Cohen's d	Standardized difference in transferred claim score between positive and negative target groups.
continuous	Spearman correlation	Rank association between transferred claim score and continuous target.
continuous	Pearson correlation	Linear association between transferred claim score and continuous target.
none or unsupported	not applicable	Transfer is not computed when no target is defined or feature matching is not available.

C.8 Stability, transfer, and overall labels

SCS assigns descriptive labels after computing the metric profile. The labels summarize the underlying metrics for interpretation. Stability and transfer labels use prespecified, configurable thresholds, with E4 evaluated under these same thresholds. The thresholds used in the experiments are summarized in Table C5, and E4 label sensitivity is reported in Appendix F.5.

Table C5: Label assignment rules used in the experiments. Labels accompany the underlying stability and transfer metrics and summarize the multidimensional profile.

Label	Condition	Meaning
insufficient stability	fewer than 5 matched ROIs or fewer than 5 matched groups	Too few matched units to evaluate stability reliably.
stable	evidence Spearman ≥ 0.70, support Jaccard ≥ 0.40, and direction retention ≥ 0.80	Evidence ranking, support set, and direction are all strongly preserved.
moderate	evidence Spearman ≥ 0.40, support Jaccard ≥ 0.20, and direction retention ≥ 0.60	The claim is partially preserved but shows nontrivial drift.
fragile	evaluated but does not meet moderate stability criteria	The claim changes substantially under the perturbation.
strong binary transfer	oriented AUROC ≥ 0.75 and $\lvert d\rvert \geq 0.50$	The transferred claim score strongly separates the target binary groups.
moderate binary transfer	oriented AUROC ≥ 0.65 and $\lvert d\rvert \geq 0.30$	The transferred claim score moderately separates the target binary groups.
weak binary transfer	binary transfer evaluated but below moderate thresholds	The claim score has limited binary transfer utility.
strong continuous transfer	$\lvert\rho_{\mathrm{Spearman}}\rvert \geq 0.35$	The transferred claim score has strong rank association with the target.
moderate continuous transfer	$\lvert\rho_{\mathrm{Spearman}}\rvert \geq 0.20$	The transferred claim score has moderate rank association with the target.
weak continuous transfer	continuous transfer evaluated but below moderate threshold	The claim score has limited continuous transfer utility.
not applicable transfer	transfer target is absent or transfer is not defined	Stability can still be evaluated, but transfer is not part of the profile.
stable and transferable	stability label is stable and transfer label is strong or moderate	The claim is preserved and useful in the target setting.
stable not transferable	stability label is stable but transfer is weak, not applicable, or insufficient	The claim is structurally preserved but does not show useful transfer.
transferable but drifted	stability label is moderate or fragile and transfer label is strong or moderate	The claim score transfers, but the supporting evidence pattern changes.
overall fragile	stability label is fragile and transfer is not strong or moderate	The claim is not reliably preserved under the perturbation.
overall insufficient	stability label is insufficient, or stability label is moderate and transfer is weak, not applicable, or insufficient	The pair lacks enough matched information or sufficient joint stability–transfer evidence for a stronger overall conclusion.

Stability and transfer can diverge. A claim can preserve its support regions without producing a useful target score, or it can transfer as a score while drifting anatomically. SCS reports the full metric profile alongside the descriptive labels to retain these distinctions.

D Empirical Design and Reproducibility Details

D.1 Dataset views and targets

Table D1 summarizes the dataset views used in the empirical evaluation. The total analysis set contains 10,370 MRI scans across ADNI, AIBL, and OASIS-1. Per-view scan counts are computed from the harmonized metadata and feature-matrix inventories. OASIS-1 is treated separately from ADNI and AIBL because it is not used with the same CN/MCI/AD harmonization; when a weak-label contrast is needed, OASIS-1 uses CDR-derived cognitive impairment.

Table D1: Dataset views used by SCS. The table separates full, baseline, and matched views because different evaluation experiments require different levels of alignment.

Dataset view	Dataset	Scan count	Targets	Role in evaluation
`ADNI`	ADNI [1]	8,637	CN/MCI/AD, MMSE, age, sex, field strength, cohort	Full ADNI view used for feature construction and claim-generation inventory.
`ADNI_baseline`	ADNI [1]	1,723	CN/MCI/AD, MMSE, age, sex, field strength, cohort	Baseline ADNI view used as the main source for volume transfer, FastSurfer drift, atlas sensitivity, and feature-family sensitivity.
`AIBL`	AIBL [3]	1,297	CN/MCI/AD-compatible flags, MMSE, age, sex, imaging protocol	Full AIBL view used for feature construction and claim-generation inventory.
`AIBL_baseline`	AIBL [3]	688	CN/MCI/AD-compatible flags, MMSE, age, sex, imaging protocol	External target for ADNI-to-AIBL volume transfer and atlas/feature-family sensitivity.
`OASIS-1`	OASIS-1 [2]	436	CDR-derived impairment, MMSE, age, sex, education	Full OASIS-1 view used for feature construction and weak-label claim-generation inventory.
`OASIS-1_baseline`	OASIS-1 [2]	436	CDR-derived impairment, MMSE, age, sex, education	External weak-label target for ADNI-to-OASIS transfer and atlas/feature-family sensitivity.
`ADNI_SPM_MATCHED`	ADNI [1]	2,892	CN/MCI/AD, MMSE, age	Matched FreeSurfer–SPM/CAT12 cortical-thickness view used for direct pipeline-drift evaluation.

D.2 Feature spaces and matched matrix sizes

Feature spaces. Table D2 lists the feature spaces used by the framework. Each feature space is defined by a feature family, processing pipeline, atlas or region system, and the dataset views in which it is available. Feature vectors are parsed into scan-by-region matrices and paired with harmonized metadata before claim extraction.

Table D2: Feature spaces used by SCS. The inventory distinguishes volume, cortical-thickness, pipeline-perturbation, and matched feature spaces.

Feature space	Family	Pipeline	Atlas or region system	Available dataset views
`fs8_cortical_volume`	cortical regional volume	FreeSurfer 8 [5]	FreeSurfer cortical anatomical regions [31, 32]	ADNI, ADNI baseline, AIBL, AIBL baseline, OASIS-1, OASIS-1 baseline
`fs8_subcortical_volume`	subcortical volume	FreeSurfer 8 [5]	FreeSurfer subcortical anatomical regions	ADNI, ADNI baseline, AIBL, AIBL baseline, OASIS-1, OASIS-1 baseline
`fs8_thickness_schaefer200_7`	cortical thickness	FreeSurfer 8 [5]	Schaefer 200 / 7-network atlas [23]	ADNI, ADNI baseline, AIBL, AIBL baseline, OASIS-1, OASIS-1 baseline, ADNI SPM-matched
`fs8_thickness_kong200_17`	cortical thickness	FreeSurfer 8 [5]	Kong 200 / 17-network atlas [24]	ADNI, ADNI baseline, AIBL, AIBL baseline, OASIS-1, OASIS-1 baseline
`fs8_thickness_yale696`	cortical thickness	FreeSurfer 8 [5]	Yale Brain Atlas [25]	ADNI, ADNI baseline, AIBL, AIBL baseline, OASIS-1, OASIS-1 baseline
`spm_thickness_schaefer200_7`	cortical thickness	SPM/CAT12 [7, 14]	Schaefer 200 / 7-network atlas [23]	ADNI SPM-matched
`spm_thickness_schaefer200_17`	cortical thickness	SPM/CAT12 [7, 14]	Schaefer 200 / 17-network mapping [23]	ADNI SPM-matched
`fastsurfer_cortical_volume`	cortical regional volume	FastSurfer [6]	FreeSurfer-compatible cortical anatomical regions	ADNI baseline
`fastsurfer_subcortical_volume`	subcortical volume	FastSurfer [6]	FreeSurfer-compatible subcortical anatomical regions	ADNI baseline

Matched matrix sizes. The ADNI SPM-matched view is used for direct cortical-thickness pipeline-drift evaluation. It contains the same 2,892 scans with aligned FreeSurfer 8 and SPM/CAT12 Schaefer thickness representations. Table D3 gives the matched matrix sizes.

Table D3: Matched matrix sizes for the ADNI FreeSurfer–SPM/CAT12 cortical-thickness comparison. All three matrices contain the same 2,892 matched scans and 202 columns.

Dataset view / feature space	Rows	Columns	Use
`ADNI_SPM_MATCHED/spm_thickness_schaefer200_7`	2,892	202	SPM/CAT12 Schaefer 7-network thickness target for pipeline-drift evaluation.
`ADNI_SPM_MATCHED/spm_thickness_schaefer200_17`	2,892	202	SPM/CAT12 Schaefer 17-network thickness target for network-resolution drift evaluation.
`ADNI_SPM_MATCHED/fs8_thickness_schaefer200_7`	2,892	202	FreeSurfer 8 Schaefer 7-network thickness source for matched pipeline comparison.

D.3 Claim Generation Inventory

D.3.1 Claim-generation settings

Claim generation is performed by expanding configured analysis settings across dataset views, feature spaces, targets, covariates, claim types, and support criteria. Each expanded setting defines one claim family. The implementation then applies the appropriate claim extractor to produce one or more claim cards from that family. The resulting inventory contains 121 claim families and 2,783 claim cards.

The claim-generation configuration has three components. First, dataset and feature-space configurations define which morphometry matrices are available. Second, claim-extraction configurations define targets, covariates, inclusion filters, claim types, and minimum sample requirements. Third, claim-generation settings define which claim levels are generated, including composite, atomic ROI, group-summary, ranking, direction-summary, latent-axis, and normative-deviation claims.

New datasets, feature spaces, targets, and claim templates can be added through configuration without changing the core extraction code.

D.3.2 Claim-family and claim-card inventory

Table D4 summarizes the claim-generation output used in the paper. The full claim-card table and claim-family table are released as structured output artifacts.

Table D4: Claim-generation summary. Each analysis setting produces one claim family, and each claim family can produce multiple claim cards at different levels of granularity.

Quantity	Value
Analysis settings	121
Claim families	121
Claim cards	2,783
Claim-card output	`data/processed/claim_cards/claim_cards.jsonl` and `claim_cards_summary.csv`
Claim-family output	`data/processed/claim_cards/claim_families.jsonl` and `claim_family_summary.csv`
Per-setting ROI statistics	`data/processed/claim_cards/per_setting/`

D.3.3 Generated claim levels and claim types

The generated claim inventory covers supervised, weak-label, unsupervised, and normative settings. Table D5 lists the claim levels and claim types included in the generated pool. Counts by dataset, feature space, target, claim level, and claim type are reported in the released claim-card summary table.

Table D5: Generated claim levels and claim types. The claim-generation inventory includes multiple granularities of scientific statements from the same underlying evidence models.

Category	Generated values
claim levels	`composite`, `atomic_roi`, `group_summary`, `ranking`, `direction_summary`, `latent_axis`, `latent_axis_association`, `normative_deviation`, `normative_atomic_roi`
claim types	`diagnosis_contrast`, `cdr_impairment_contrast`, `mmse_association`, `age_association`, `continuous_association`, `demographic_contrast`, `acquisition_contrast`, `feature_ranking`, `group_enrichment`, `direction_summary`, `latent_morphometry_axis`, `latent_axis_association`, `normative_deviation`, `normative_group_deviation`

D.3.4 Claim-generation outputs used downstream

Claim-generation outputs are used in review and evaluation. First, claim cards provide the candidate pool for expert and LLM review. Second, claim families provide the source and target objects used in stability and transfer evaluation. Third, per-setting ROI statistics provide the evidence vectors, support regions, and anatomical annotations used to compute exact or group-level stability metrics.

SCS evaluates claims at the family level and retains the individual claim cards. For example, a diagnosis-contrast family can contain a composite claim, atomic ROI claims, group-summary claims, a feature-ranking claim, and a direction-summary claim. The family-level evidence is then compared across settings, while selected claim cards are exported for human and LLM review.

D.4 Evaluation Experiment Details

D.4.1 Evaluation-pair construction

SCS evaluates stability and transfer through ordered source–target pairs. Each evaluation pair is defined as

$$\pi = (s_{\mathrm{src}}, s_{\mathrm{tgt}}, \delta, \mu, \tau_{\mathrm{src}}, \tau_{\mathrm{tgt}}), \tag{39}$$

where s_{src} is the source setting, s_{tgt} is the target setting, δ is the perturbation type, μ is the ROI matching strategy, and τ_{src} and τ_{tgt} are the source and target claim types. The evaluation-pair configuration specifies the source and target dataset views, feature spaces, claim types, targets, matching strategy, and transfer target when applicable. The expanded evaluation table contains one row per source–target claim-family comparison.

Evaluation requires both source and target claim families to be available. For each pair, the evaluator loads the corresponding ROI-level statistics, claim-family metadata, support regions, and feature matrices when transfer scoring is required. Pairs are evaluated using exact ROI matching when source and target feature spaces share comparable ROI names. Pairs are evaluated using group-level matching when source and target feature spaces do not share direct ROI correspondence. Pairs with no valid matched features, missing claim families, or missing ROI statistics are marked as skipped or insufficient.

In the reported experiments, 89 evaluation pairs were expanded and all 89 were evaluated. No configured pair was skipped and no pair produced a processing error.

D.4.2 Implemented evaluation experiments

Table D6 gives the full E1–E6 experiment inventory, including source–target settings, matching strategies, and target mappings.

Table D6: Full E1–E6 evaluation experiment inventory. Each row summarizes one implemented perturbation family and the claim-family pairs generated from it.

ID	Perturbation	Source setting	Target setting	Feature comparison	Targets and pairs
E1	dataset transfer	ADNI baseline	AIBL baseline	FreeSurfer 8 cortical regional volume and subcortical volume, exact ROI matching	CN–AD, CN–MCI, MMSE, and age claims are transferred from ADNI to AIBL across two volume feature spaces, giving 8 evaluated pairs.
E2	weak-label dataset transfer	ADNI baseline	OASIS-1 baseline	FreeSurfer 8 cortical regional volume and subcortical volume, exact ROI matching	MMSE and age claims are transferred directly, and ADNI CN–AD claims are evaluated against OASIS-1 CDR-derived impairment, giving 6 evaluated pairs.
E3	cortical-thickness pipeline drift	ADNI SPM-matched	ADNI SPM-matched	FreeSurfer 8 Schaefer thickness versus SPM/CAT12 Schaefer thickness, with exact ROI or network-level matching depending on comparison	CN–AD, CN–MCI, MMSE, and age claims are compared across three matched thickness comparisons, giving 12 evaluated pairs.
E4	volume pipeline drift	ADNI baseline	ADNI baseline	FreeSurfer 8 cortical/subcortical volume versus FastSurfer cortical/subcortical volume, exact ROI matching	CN–AD, CN–MCI, MMSE, and age claims are compared across two volume feature spaces, giving 8 evaluated pairs.
E5	atlas sensitivity	ADNI baseline, AIBL baseline, OASIS-1 baseline	same dataset view	FreeSurfer 8 Schaefer, Kong, and Yale cortical-thickness atlases, group-level matching	ADNI/AIBL use CN–AD, CN–MCI, MMSE, and age; OASIS-1 uses CDR-derived impairment, MMSE, and age; three atlas-pair comparisons give 33 evaluated pairs.
E6	feature-family sensitivity	ADNI baseline, AIBL baseline, OASIS-1 baseline	same dataset view	FreeSurfer 8 cortical/subcortical volumes versus FreeSurfer 8 Schaefer cortical thickness, group-level matching	ADNI/AIBL use CN–AD, CN–MCI, MMSE, and age; OASIS-1 uses CDR-derived impairment, MMSE, and age; two volume-to-thickness comparisons give 22 evaluated pairs.

D.4.3 Matching strategies

The evaluation configuration supports three matching regimes. The first is exact ROI matching, which is used when source and target ROI-statistic tables share comparable feature names. The second is group-level matching, which is used when exact ROI names do not correspond but both feature spaces can be mapped to a shared anatomical or network-level grouping. The third, no direct ROI matching, retains metadata for comparisons whose feature spaces do not support stability computation.

Exact ROI matching defines the shared feature set as

$$\mathcal{J}_{st} = \mathcal{R}_{s_{\mathrm{src}}} \cap \mathcal{R}_{s_{\mathrm{tgt}}}, \tag{40}$$

where $\mathcal{R}_{s_{\mathrm{src}}}$ and $\mathcal{R}_{s_{\mathrm{tgt}}}$ are the source and target ROI-name sets. Evidence, support, direction, magnitude, and transfer metrics are computed over the matched ROI set. Exact matching is used for E1, E2, E4, and the direct Schaefer-to-Schaefer comparisons in E3.

Group-level matching defines a shared group space using an anatomical or network-level mapping. The main group strategies are `network_7_like`, `lobe`, and `group_summary`. The fallback order prioritizes the requested grouping variable and then uses available anatomical annotations such as network, ROI group, anatomical group, lobe, or network-like group. Group-level matching is used for atlas sensitivity and feature-family sensitivity, where one-to-one ROI correspondence is not available or would be misleading.

Table D7 summarizes the matching strategies used in the experiments.

Table D7: Matching strategies used in E1–E6. Exact matching is used when ROI names are comparable, while group-level matching supports atlas and feature-family comparisons.

Strategy	Shared unit	Used in	Purpose
`exact_roi`	matched ROI names	E1, E2, E4, direct E3 comparisons	Compares evidence and support over the same feature columns or compatible anatomical names.
`network_7_like`	coarse network-like group	E3 and E5 where applicable	Compares atlas-specific cortical-thickness evidence at a shared network-like resolution.
`lobe`	cortical lobe or anatomical lobe group	E5 and E6 where applicable	Provides a coarse anatomical comparison when atlas parcels or feature families differ.
`group_summary`	ROI group or anatomical group	E6 and fallback grouped comparisons	Supports comparison between volumes and cortical thickness or between non-identical region systems.
`no_direct_roi_match`	none	not used in the main evaluated pairs	Records comparisons for which stability cannot be computed from the available feature spaces.

D.4.4 Target mappings

Evaluation pairs may compare the same target across datasets or compare related but non-identical targets. For direct transfer, the source and target represent the same scientific target, such as MMSE or age. For weak-label transfer, the target definitions differ but are scientifically related. The main weak-label case is ADNI CN–AD transferred to OASIS-1 CDR-derived impairment.

Table D8 summarizes the target mappings used across E1–E6.

Table D8: Target mappings used in evaluation experiments. OASIS-1 is treated as a weak-label setting when CDR-derived impairment is used instead of CN/MCI/AD diagnosis.

Mapping type	Source target	Target target	Used in
diagnosis transfer	ADNI CN–AD	AIBL CN–AD	E1
diagnosis transfer	ADNI CN–MCI	AIBL CN–MCI	E1
weak-label transfer	ADNI CN–AD	OASIS-1 CDR-derived impairment	E2
continuous transfer	ADNI MMSE	AIBL or OASIS-1 MMSE	E1, E2
continuous transfer	ADNI age	AIBL or OASIS-1 age	E1, E2
within-view target preservation	CN–AD, CN–MCI, MMSE, age	same target in same dataset view	E3, E4, E5, E6 for ADNI and AIBL
within-view weak-label preservation	CDR-derived impairment, MMSE, age	same target in OASIS-1 baseline	E5, E6 for OASIS-1

For binary contrasts, the positive and negative labels are defined in the claim-extraction setting. For continuous targets, the target direction is retained from the source claim and the transferred claim score is evaluated by correlation with the target in the target setting. For weak-label transfer, the target measures transfer to a related clinical construct; the resulting scores are interpreted within the target label definition.

D.4.5 Skipped-pair and error policy

The evaluator records every configured pair as evaluated, skipped, or error. A pair is marked as skipped when the source claim family is missing, the target claim family is missing, the source or target ROI statistics are missing, or no matched features are available under the configured matching strategy. A pair is marked as insufficient when the comparison can be constructed but does not have enough matched units for reliable stability estimation. A pair is marked as error only when an unexpected exception occurs during loading, matching, metric computation, transfer scoring, or label assignment.

All skipped and error cases are written to output tables with the reason. The output tables therefore record coverage of the observed multiverse. In the experiments reported in the main paper, the expanded evaluation configuration contained 89 pairs, all 89 pairs were evaluated, and no pair was skipped or marked as an error.

D.5 Implementation architecture and configuration

D.5.1 Framework architecture

SCS is implemented as a configuration-driven Python package, `scs_morph`. The package follows the scientific workflow, with separate modules for each processing and evaluation stage. Table D9 summarizes the main implementation layers used in the experiments.

Table D9: `scs_morph` implementation architecture. The framework separates data preparation, claim extraction, evaluation, and review so that each stage can be reproduced and audited independently.

Layer	Main modules or scripts	Role
data loading and audit	`scs_morph.data`	Loads raw tabular inputs, audits column encodings, validates vectorized feature columns, and constructs dataset-level audit reports.
metadata harmonization	`scs_morph.harmonization`	Harmonizes dataset-specific identifiers, demographic variables, diagnosis labels, cognitive scores, imaging metadata, and baseline or matched views.
feature matrix construction	`scs_morph.features`	Converts vector-valued feature columns into scan-by-region matrices and aligned metadata tables for each dataset view and feature space.
analysis preparation	`scs_morph.analysis`	Builds analysis tables, applies inclusion rules, encodes covariates, standardizes features or targets when configured, and prepares design matrices.
claim extraction	`scs_morph.claims`	Fits claim-specific evidence models, applies support criteria, annotates regions, renders templates, and serializes claim cards and claim families.
claim evaluation	`scs_morph.evaluation`	Constructs evaluation pairs, matches exact or grouped feature spaces, computes stability and transfer metrics, and assigns stability, transfer, and overall labels.
expert and LLM review	`scs_morph.expert`	Selects review claims, exports review sheets, imports scored reviews, computes agreement summaries, and assigns expert-review decisions.

D.5.2 Configuration files

The workflow is controlled by YAML configuration files. The files specify dataset views, feature spaces, claim-generation rules, evaluation pairs, thresholds, and review settings independently of the code. Table D10 lists the configuration files used by the experiments.

Table D10: Configuration files used by the SCS workflow. Configuration files define the observed multiverse, claim-generation settings, evaluation pairs, thresholds, and review protocol.

Configuration file	Purpose
`paths.yaml`	Defines raw, interim, processed, and output directories.
`datasets.yaml`	Defines dataset input files and dataset identifiers.
`feature_spaces.yaml`	Defines feature-space names, source columns, feature-name files, feature families, pipelines, and atlases.
`claim_generation.yaml`	Defines claim-generation options, feature-family labels, and support settings.
`claim_extraction_all_main.yaml`	Defines claim-extraction settings to generate claim families and claim cards.
`claim_extraction_examples.yaml`	Defines smaller example settings for inspection, debugging, or documentation.
`claim_templates.yaml`	Defines structured natural-language templates for claim text generation.
`evaluation_pairs.yaml`	Defines source–target evaluation pairs, perturbation types, matching strategies, targets, and transfer settings.
`evaluation_thresholds.yaml`	Defines thresholds for stability labels, transfer labels, and overall labels.
`expert_review.yaml`	Defines expert-review selection, reviewer groups, scoring columns, categorical labels, and decision rules.

The scientific workflow layers regenerate claim cards and evaluation metrics. Manuscript-output configurations document the additional steps used to prepare paper outputs.

D.6 Output artifacts and reproducibility workflow

D.6.1 Major output artifacts

The framework writes intermediate and final artifacts to structured directories. Table D11 lists the artifacts produced by the workflow. Some artifacts, such as harmonized clinical tables, scan-by-region feature matrices, and participant-level claim scores, are generated locally but are not redistributed because they contain participant-level or derived participant-level information. Released artifacts are restricted to code, configuration files, documentation, and aggregate summaries permitted by the relevant data-use agreements.

Table D11: Major output artifacts produced by the SCS workflow. The outputs support auditing of claim generation, evaluation, and expert review.

Artifact	Content
Harmonized datasets and feature matrices	
`harmonized/`	Harmonized dataset views, harmonization reports, and warnings.
`feature_matrices/`	Scan-by-region feature matrices and aligned metadata tables for each dataset view and feature space.
Claim-card outputs: `data/processed/claim_cards/`	
`claim_cards.jsonl`	Serialized claim cards generated from all analysis settings.
`claim_families.jsonl`	Serialized claim families linking all claim cards extracted from the same setting.
`claim_cards_summary.csv`	Tabular summary of claim cards for counting, filtering, and review selection.
`claim_family_summary.csv`	Tabular summary of claim families and source settings.
`per_setting/`	ROI-level statistics, claim-family metadata, and setting-specific artifacts.
Evaluation outputs: `data/processed/evaluations/`	
`evaluation_pairs_expanded.csv`	Expanded source–target evaluation-pair table.
`evaluation_results.csv`	One-row-per-pair evaluation results with labels and metric summaries.
`evidence_stability.csv`	Evidence-concordance metrics for evaluated pairs.
`support_stability.csv`	Support-overlap metrics for evaluated pairs.
`direction_stability.csv`	Direction-retention and sign-flip metrics.
`magnitude_stability.csv`	Magnitude-drift metrics.
`group_stability.csv`	Group-level stability metrics for non-identical feature spaces.
`transfer_metrics.csv`	Claim-score transfer metrics for binary and continuous targets.
Expert-review outputs: `data/processed/expert_review/`	
`selected_main_expert_claims.csv`	Main human expert review set.
`selected_llm_extended_claims.csv`	Extended LLM review set.
`selected_reserve_claims.csv`	Reserve claim set for replacement or secondary review.
`analyzed/`	Imported review scores, agreement summaries, human–LLM comparisons, and expert-decision tables.

D.6.2 Reproducibility workflow

The implementation is released as a reproducible code repository. The repository contains the `scs_morph` package, configuration files, and command-line scripts used to regenerate the harmonized data views, feature matrices, claim cards, evaluation results, and review summaries. Participant-level MRI data, clinical tables, scan identifiers, scan-level feature matrices, and participant-level derived outputs are not redistributed.

The end-to-end workflow is script-based. Each step reads the configured inputs, writes explicit output artifacts, and can be rerun independently after authorized source data are placed under the configured raw-data directory. Table D12 summarizes the recommended execution order.

Workflow stages can be rerun separately. For example, after claim cards have been generated, new evaluation pairs can be added by editing `configs/evaluation_pairs.yaml` and rerunning only `scripts/evaluate_claims.py`. Similarly, new claim-review batches can be created by editing `configs/expert_review.yaml` and rerunning the expert-review export script. Additional datasets, pipelines, atlases, feature families, and review protocols can be incorporated through these modules while retaining the core SCS definitions.

Table D12: Reproducibility workflow. The SCS evaluation can be regenerated by running the scripts in order after obtaining authorized access to the source datasets.

Step	Script	Purpose and output
1	`audit_data.py`	Audits raw input tables, summarizes column encodings, missingness, and frequent values, and writes raw-data audit reports.
2	`parse_data.py`	Parses raw vector-valued morphometry fields and checks whether feature vectors can be interpreted consistently.
3	`harmonize_data.py`	Harmonizes dataset-specific identifiers, visits, demographics, diagnoses, cognitive scores, imaging metadata, and dataset views.
4	`build_feature_matrices.py`	Converts harmonized vector-valued feature columns into scan-by-region matrices and aligned metadata tables for each configured feature space.
5	`extract_claims.py`	Expands claim-generation settings, fits claim-extraction models, writes ROI-level statistics, and serializes claim families and claim cards.
6	`evaluate_claims.py`	Expands source–target evaluation pairs, performs exact or group-level matching, computes stability and transfer metrics, and assigns descriptive labels.
7	`report_evaluation_results.py`	Summarizes evaluation outputs into experiment-level and metric-specific reports for inspection and manuscript analysis.
8	`export_expert_review.py`	Selects claim cards for human expert review, LLM review, and reserve review, and exports review sheets.
9	`import_expert_reviews.py`	Imports returned review sheets and converts reviewer scores, categorical decisions, corrected text, and comments into analysis-ready tables.
10	`analyze_expert_reviews.py`	Computes claim-level review summaries, agreement metrics, human–LLM agreement, high-disagreement flags, and expert-review decisions.

E Expert and LLM Review Protocol

E.1 Claim selection for review

SCS uses structured review to assess whether generated claim cards are scientifically plausible, clearly written, faithful to their evidence, and useful for interpretation. The review protocol assesses claim cards containing group-level scientific statements derived from morphometry evidence. Reviewers are instructed to evaluate these statements at the group level.

The review set is sampled from the generated claim-card pool to cover multiple datasets, feature spaces, claim levels, and claim types. The main human expert review set contains 60 claim cards, each scored by all three human reviewers. The extended LLM review set contains these 60 claims plus 60 additional claims, giving 120 LLM-reviewed cards in total. A reserve set of 20 additional claims was prepared for possible replacements but was not used in the reported review. The human review panel includes three domain perspectives: neuroscience, geriatrics, and radiology. The LLM judges provide auxiliary review scores, with agreement to human reviewers assessed empirically. No participant-level MRI data, scan-level feature vectors, clinical rows, subject identifiers, or derived participant-level data are provided to LLM judges; review inputs contain only aggregate claim-card text and non-identifying statistical summaries.

The selected human review set is stratified across claim levels and claim types. It includes composite, atomic ROI, group-summary, ranking, latent-axis, latent-axis association, normative-deviation, and normative atomic ROI claims. It also includes diagnosis contrasts, MMSE associations, group-enrichment claims, age associations, feature-ranking claims, CDR-derived impairment claims, latent

morphometry-axis claims, and normative-deviation claims. Stratification permits assessment of claim-card validity across claim granularities, targets, datasets, and feature families.

Each reviewed claim card is exported as one row in a review table. The row contains the generated claim text, source setting, feature family, target, anatomical fields, support information, and statistical evidence required for review. Reviewers score the claim as written, but may optionally provide a corrected claim text and short comments when the wording is misleading, too broad, insufficiently cautious, or anatomically unclear.

E.2 Review dimensions and scoring rubric

Each reviewer $r \in \mathcal{R}$ scores each claim c on five ordinal dimensions. For dimension d, the score is

$$a_{r,d}(c) \in \{1, 2, 3, 4, 5\}. \tag{41}$$

The dimensions are plausibility, clarity, faithfulness, usefulness, and reviewer confidence. Table E1 defines the question associated with each dimension.

Table E1: Expert and LLM review dimensions. Each dimension targets a different aspect of claim-card validity.

Dimension	Question
plausibility	Is the claim biologically or clinically plausible given current neuroanatomical and clinical knowledge?
clarity	Is the claim understandable, specific, and well phrased?
faithfulness	Does the claim match the evidence shown in the row, including ROI, direction, statistics, and feature type?
usefulness	Would the claim help interpret MRI morphometry results?
reviewer confidence	How confident is the reviewer in their judgment of this claim?

All dimensions use the same numerical scoring rubric. A score of 1 indicates a poor rating on the corresponding dimension, while a score of 5 indicates a strong rating. Table E2 gives the common interpretation of the 1–5 scale.

Table E2: Review score rubric. The same 1–5 scale is used for plausibility, clarity, faithfulness, usefulness, and confidence, with dimension-specific wording provided in the review sheet.

Score	Meaning
1	Very low rating, such as implausible, unclear, unfaithful, not useful, or very low confidence.
2	Weak rating, such as questionable, somewhat unclear, weakly faithful, low usefulness, or low confidence.
3	Moderate rating, such as plausible but uncertain, understandable but improvable, mostly faithful with caution, moderately useful, or moderate confidence.
4	Strong rating, such as plausible, clear, faithful, useful, or high confidence.
5	Very strong rating, such as highly plausible, very clear, very faithful, highly useful, or very high confidence.

Reviewers also assign two categorical judgments. The caution level takes values `stable`, `tentative`, `fragile`, `reject`, or `not_sure`. The recommended action takes values `accept`, `minor_revise`, `major_revise`, `reject`, or `not_sure`. The review instructions reward cautious, faithful wording and penalize unsupported causal, diagnostic, or mechanistic claims. Reviewers are encouraged to use `not_sure` and lower confidence when a claim falls outside their expertise.

E.3 Agreement and decision rules

SCS summarizes expert review at the claim level and at the dimension level. For each claim c and dimension d, the mean human score is

$$\bar{a}_d^{\text{human}}(c) = \frac{1}{|\mathcal{R}_{\text{human}}(c)|} \sum_{r \in \mathcal{R}_{\text{human}}(c)} a_{r,d}(c), \tag{42}$$

where $\mathcal{R}_{\text{human}}(c)$ denotes the human reviewers who scored claim c. The same summary can be computed for LLM judges or for all reviewers.

Reviewer agreement is summarized using pairwise correlations and intraclass correlation coefficients. For each dimension d, SCS computes pairwise Pearson and Spearman correlations between reviewer score vectors on overlapping claims. Let $\mathbf{a}_{r,d}$ and $\mathbf{a}_{r',d}$ denote two reviewers' score vectors for dimension d over shared claims. The pairwise Spearman agreement is

$$\rho_{rr',d} = \rho_{\text{Spearman}}\left(\mathbf{a}_{r,d}, \mathbf{a}_{r',d}\right). \tag{43}$$

For human experts, SCS also reports ICC-style agreement. Given a claim-by-reviewer score matrix for one dimension, let n be the number of claims, k be the number of reviewers, MS_R be the mean square for rows, MS_C the mean square for columns, and MS_E the residual mean square. The single-reviewer ICC is

$$\text{ICC}(2,1) = \frac{MS_R - MS_E}{MS_R + (k-1)MS_E + \frac{k}{n}(MS_C - MS_E)}. \tag{44}$$

The average-reviewer ICC is

$$\text{ICC}(2,k) = \frac{MS_R - MS_E}{MS_R + \frac{1}{n}(MS_C - MS_E)}. \tag{45}$$

These agreement summaries follow standard intraclass-correlation formulations for reviewer reliability [37, 38].

Human–LLM agreement is evaluated by comparing each LLM judge to the mean human score. For LLM judge ℓ and dimension d, SCS computes

$$\rho_{\ell,d}^{\text{human}} = \rho_{\text{Spearman}}\left(\mathbf{a}_{\ell,d}, \bar{\mathbf{a}}_{d}^{\text{human}}\right), \tag{46}$$

along with mean absolute difference and signed bias. These measures quantify the agreement of auxiliary LLM judges with human experts.

SCS also assigns claim-level review decisions. A high-disagreement flag is assigned when the human standard deviation for plausibility or faithfulness exceeds a configured threshold. In the experiments, this threshold is 1.25. Expert rejection is assigned when mean human plausibility or mean human faithfulness is at most 2.5, or when the majority recommended action is `reject`. Expert acceptance or minor revision is assigned when mean human plausibility is at least 4, mean human faithfulness is at least 4, and mean human usefulness is at least 3.5, with the final label depending on the majority recommended action. Claims that do not satisfy acceptance or rejection criteria are assigned major revision, expert disagreement, or insufficient review depending on review coverage and disagreement status. Table E3 summarizes these decision rules.

Table E3: Expert-review decision rules. Decision labels summarize human review outcomes while preserving the underlying scores, comments, caution levels, and recommended actions.

Decision	**Rule**
insufficient reviews	Fewer than two human reviewers scored the claim.
expert disagreement	Human-review standard deviation for plausibility or faithfulness is at least 1.25.
expert rejected	Mean human plausibility ≤ 2.5, mean human faithfulness ≤ 2.5, or majority action is `reject`.
expert accepted	Mean human plausibility ≥ 4, mean human faithfulness ≥ 4, mean human usefulness ≥ 3.5, and majority action is `accept`.
expert minor revision	Mean human plausibility ≥ 4, mean human faithfulness ≥ 4, mean human usefulness ≥ 3.5, but majority action is not `accept`.
expert major revision	Claim is sufficiently reviewed and not rejected, accepted, minor-revised, or high-disagreement.

Alongside statistical evaluation, the review protocol assesses whether generated claims are scientifically interpretable, appropriately cautious, and faithful to the displayed evidence.

E.4 LLM auxiliary-review prompt

LLM judges use the same review dimensions and categorical labels as human reviewers. The full auxiliary-review prompt is stored in `configs/llm_expert_review_prompt.md`. The prompt instructs LLM judges to act as conservative scientific reviewers, score group-level MRI morphometry claims, avoid patient-level diagnosis, avoid unsupported mechanisms, and return only a completed CSV-compatible review table. The following excerpt documents the main review constraints.

> You are acting as an expert scientific reviewer for a research project titled: "A Scientific Claim Stability Framework for Evaluating MRI Morphometry Pipelines".
>
> ...
>
> The claims are group-level scientific statements, not individual patient diagnoses.
>
> Important:
>
> - Do not infer patient-level diagnosis.
> - Do not add unsupported biological mechanisms.
> - Do not reward claims just because they sound familiar.
> - Focus on whether the claim is plausible, clear, faithful to the provided evidence, and useful.
> - Use conservative scientific judgment.
> - Do not change `review_id` or `claim_id`.
>
> ...
>
> Please fill in the following scoring columns: `plausibility_1_5`, `clarity_1_5`, `faithfulness_1_5`, `usefulness_1_5`, and `reviewer_confidence_1_5`.
>
> ...
>
> Choose exactly one `caution_level`: `stable`, `tentative`, `fragile`, `reject`, or `not_sure`.
>
> Choose exactly one `recommended_action`: `accept`, `minor_revise`, `major_revise`, `reject`, or `not_sure`.
>
> ...
>
> Return only the completed CSV table. Preserve all original rows. Preserve `review_id` and `claim_id` exactly. Use numeric values 1–5 for score columns. Use only the allowed categorical values.

The prompt uses the same scoring structure as the human review sheet and requires CSV output, unchanged identifiers, and interpretations limited to the supplied evidence.

F Additional Results

F.1 Detailed descriptions of the main figures

F.1.1 Figure 1: Scientific Claim Stability framework

Panel A: Conventional evaluation. The conventional workflow converts structural MRI scans into regional morphometry tables and evaluates pipeline outputs using measurement comparisons, agreement statistics, or predictive performance.

Panel B: Claim-centric evaluation. SCS extracts scientific statements from the same ROI tables and evaluates the evidence and anatomical support for each statement. The diagram shows contrast, association, ranking or group, and latent or normative claims, alongside the complementary dimensions of stability, transfer, drift, and review.

Panel C: Reusable `scs_morph` framework. The workflow moves through data loading and audit, harmonization, feature construction, claim extraction and evaluation, systematic reporting, expert review, and review analysis. The feedback connection shows how reviewer assessments can be used to inspect and refine generated claims.

Panel D: Perturbation axes. The framework compares claims across changes in dataset, processing pipeline, atlas or ROI system, feature family, and scientific target. The icons represent the differences captured by these axes, including sample composition, processing choices, regional representations, morphometric measurements, and target definitions.

Panel E: Claim-card representation. The schematic links a structured claim object, with its type, evidence vector, support regions, weights, and metadata, to a human-readable card. The support, stability, transfer, drift, and expert-review indicators show how evaluation results can be presented alongside a claim; the statuses shown are illustrative.

F.1.2 Figure 2: Claim landscape and evaluation coverage

Panel A: Claim-generation flow. The flow diagram traces the 2,783 generated claim cards from dataset views through cortical-thickness or volume features, grouped scientific targets, and claim levels. Node sizes and connecting bands show how the cards are distributed across these categories, and colors identify the grouped claim types.

Panel B: Claim-card landscape by level. Each horizontal bar shows the percentage composition of a claim level across diagnosis, age, cognition, confounder, and unsupervised targets. Markers on the right give the number of cards and claim families at each level, distinguishing the large number of atomic ROI cards from the smaller groups of summary claims.

Panel C: Dataset–feature availability. The left plot shows MRI data-point counts by dataset view, together with the available feature-space counts and total feature dimensions. The adjacent matrix reports the number of claim cards generated for each dataset–feature combination; dashes indicate feature spaces that were available but not used to generate claims in the reported evaluation.

Panel D: E1–E6 evaluation map. Each row reports the number of evaluated claim-family pairs in an experiment, with marker color indicating the median support Jaccard. The E2 experiment is divided into direct cross-dataset transfer (E2a) and CDR-derived weak-label transfer (E2b). The remaining rows cover E1 dataset transfer, E3–E4 pipeline drift, E5 atlas sensitivity, and E6 feature-family sensitivity.

Panel E: Claim-family stability. The matrix cross-tabulates stability and transfer labels across the 89 evaluated pairs, with marginal totals and tags for the resulting overall-label rules. It distinguishes stable, moderate, fragile, and insufficient stability from strong, moderate, weak, and insufficient transfer, according to the classification rules in Appendix C.8.

F.1.3 Figure 3: Stability, transfer, and fragile support

Panel A: Evidence concordance versus support overlap. Each evaluated pair is positioned according to the Spearman concordance of its signed evidence vectors and the Jaccard overlap of its supporting regions or matched groups. Point colors indicate the overall evaluation labels, dashed boundaries mark the stability thresholds, and crosses identify pairs for which both plotted metrics are unavailable. This view distinguishes preservation of the global evidence ranking from preservation of the regions supporting a claim.

Panel B: Transfer metrics. The upper plot shows oriented AUROC for binary-target transfer, and the lower plot shows the absolute Spearman correlation for continuous-target transfer. Both metrics are plotted against evidence concordance, with point colors indicating the overall evaluation labels and horizontal guides marking the displayed transfer-strength regimes.

Panel C: Pipeline drift collapse. The eight E4 FreeSurfer 8–FastSurfer volume pairs are connected between their evidence Spearman and support Jaccard values. Pair identifiers and target types are shown in the inset. Most pairs retain high evidence concordance but have much lower support overlap, illustrating the fragility observed under the primary E4 specification.

Panel D: Fragility gap cases. Selected comparisons are ordered by the difference between evidence concordance and support overlap, calculated as evidence Spearman minus support Jaccard. The adjacent markers show evidence concordance, support Jaccard, and direction retention for each case. Large positive gaps indicate relatively preserved evidence ranking alongside low anatomical-support overlap, whereas negative gaps indicate the reverse pattern.

Panel E: Representative claim cards. Six card designs illustrate composite, atomic ROI, group-summary, ranking, latent-axis, and normative-deviation claims. Each shows how a statement can be accompanied by its source setting, anatomical support, evidence, and evaluation context. These cards illustrate the claim formats; not every displayed claim type is included in the E1–E6 pairwise evaluations.

F.1.4 Figure 4: Anatomical, representation, and review summaries

Panel A: Lobe marker brains. The matrix counts supported claim cards by anatomical group and target category, including diagnosis or impairment, age association, cognition, and confounder associations. The adjacent brain schematics show the corresponding broad support patterns for

each target category. Counts refer to generated claims assigned to anatomical categories, not to independent participants or unique anatomical regions.

Panel B: Network and hemisphere summary. Mirrored horizontal bars show supported cortical-thickness claim counts by network, hemisphere, and target category. The left and right halves distinguish the hemispheres, while the colored segments show contributions from diagnosis, age, cognition, and confounder claims.

Panel C: Representation-sensitivity matrix. Rows compare the atlas and feature-family representations used in E5 and E6, while columns distinguish diagnosis or impairment, age association, cognition, and confounder targets. Cell values report weighted mean evidence Spearman and the number of evaluated pairs; color strips summarize the corresponding stability labels. Cells marked `n/a` indicate that no applicable evidence-concordance summary was available for that comparison.

Panel D: Reviewer scores. The radar plot compares mean plausibility, clarity, faithfulness, usefulness, and confidence scores from the geriatrics, neuroscience, and radiology experts and the two auxiliary LLM reviewers. All dimensions use the five-point review scale defined in Appendix E.2.

Panel E: Human–LLM agreement. For each review dimension, the plot shows the Spearman correlation between an LLM reviewer's scores and the mean score assigned by the human experts across the 60 claims reviewed by both. The annotations b indicate signed scoring bias relative to the human mean, with positive values denoting higher LLM ratings.

Panel F: Human-review decision matrix. The matrix cross-tabulates final human-review decisions by grouped claim type and level for the 60 human-reviewed cards. Cell values give the number of claims in each outcome: accepted, minor revision, major revision, rejected, or expert disagreement.

F.2 Claim landscape inventory

Tables F1–F3 provide the numeric inventory underlying Figure 2. The tables summarize the generated claim-card landscape by claim level, grouped claim type, and dataset–feature coverage. The grouped claim-type labels summarize the formal taxonomy in Appendix B.4 for visualization and reporting.

Table F1: Claim-card counts by claim level. Atomic ROI claims preserve fine-grained regional evidence, while higher-level claim types provide interpretable summaries for review and transfer.

Claim level	Claim cards	Proportion
`atomic_roi`	2,081	0.748
`group_summary`	259	0.093
`latent_axis_association`	97	0.035
`direction_summary`	84	0.030
`composite`	84	0.030
`ranking`	84	0.030
`normative_atomic_roi`	56	0.020
`latent_axis`	33	0.012
`normative_deviation`	5	0.002
Total	**2,783**	**1.000**

Table F2: Claim-card counts by grouped claim type. Grouped claim types are used for compact visualization of the claim landscape.

Grouped claim type	Claim cards	Proportion
`diagnosis_or_impairment`	1,184	0.425
`age`	708	0.254
`cognition`	597	0.215
`demographic_or_acquisition`	261	0.094
`latent_or_normative`	33	0.012
Total	**2,783**	**1.000**

Table F3: Dataset and feature coverage used for claim generation. MRI rows count non-missing feature rows in the corresponding dataset view and feature group; claim-card counts are zero for resource views used for coverage accounting but not directly expanded into claim families.

Dataset view	Feature group	MRI rows	Feature spaces	Feature dimensions	Claim cards
ADNI	FS8 thickness	8,637	3	554	–
ADNI	FS8 volume	8,455	2	132	–
ADNI	FastSurfer volume	3,503	2	132	–
ADNI_SPM_MATCHED	FS8 thickness	2,892	3	554	148
ADNI_SPM_MATCHED	FS8 volume	2,892	2	132	–
ADNI_SPM_MATCHED	FastSurfer volume	1,353	2	132	–
ADNI_SPM_MATCHED	SPM/CAT12 thickness	2,892	2	404	300
ADNI_baseline	FS8 thickness	1,723	3	554	505
ADNI_baseline	FS8 volume	1,723	2	132	405
ADNI_baseline	FastSurfer volume	798	2	132	201
AIBL	FS8 thickness	1,297	3	554	–
AIBL	FS8 volume	1,296	2	132	–
AIBL_baseline	FS8 thickness	688	3	554	399
AIBL_baseline	FS8 volume	685	2	132	261
OASIS-1	FS8 thickness	436	3	554	–
OASIS-1	FS8 volume	436	2	132	–
OASIS-1_baseline	FS8 thickness	436	3	554	365
OASIS-1_baseline	FS8 volume	436	2	132	199

F.3 Experiment-level evaluation summary

Table F4 expands the compact experiment-level summary in Table 1. It reports all label distributions and median metrics used to interpret the six perturbation experiments. Direction retention is computed over the source support set.

Table F4: Full E1–E6 evaluation summary. EC denotes median evidence Spearman, SO denotes median support Jaccard, and DR denotes median direction retention over source-support regions.

Experiment	Eval.	Skip	Stability counts	Transfer counts	Overall counts	EC	SO	DR	AUROC	$\lvert\rho\rvert$
E1_ADNI_TO_AIBL_VOLUME_TRANSFER	8	0	stable: 6; moderate: 2	strong: 4; moderate: 3; weak: 1	stable-transferable: 6; transferable-drifted: 1; insufficient: 1	0.891	0.633	1.000	0.718	0.352
E2_ADNI_TO_OASIS_VOLUME_TRANSFER	6	0	stable: 5; moderate: 1	strong: 5; moderate: 1	stable-transferable: 5; transferable-drifted: 1	0.842	0.703	1.000	0.741	0.634
E3_FS8_VS_SPM_THICKNESS_DRIFT	12	0	stable: 10; moderate: 2	insufficient: 8; strong: 2; moderate: 1; weak: 1	stable-not-transferable: 8; stable-transferable: 2; transferable-drifted: 1; insufficient: 1	0.875	1.000	1.000	0.704	0.363
E4_FS8_VS_FASTSURFER_VOLUME_DRIFT	8	0	fragile: 8	weak: 8	fragile: 8	0.903	0.098	0.546	0.561	0.137
E5_FS8_THICKNESS_ATLAS_SENSITIVITY	33	0	stable: 16; moderate: 7; fragile: 10	insufficient: 33	stable-not-transferable: 16; fragile: 10; insufficient: 7	0.700	1.000	1.000	–	–
E6_FS8_VOLUME_VS_THICKNESS_SENSITIVITY	22	0	stable: 4; moderate: 1; fragile: 6; insufficient: 11	insufficient: 22	insufficient: 12; fragile: 6; stable-not-transferable: 4	0.500	1.000	1.000	–	–

The high median DR for E5 and E6 describes direction retention in the grouped anatomical spaces frequently used by these comparisons. Preservation of fine-grained ROI support requires ROI-level evaluation.

F.4 High-concordance fragile-support cases

Tables F5–F6 provide the pair-level metrics underlying Figure 3. For readability, pair identifiers are shortened to the experiment-local suffix; full `pair_id` values are released in `figure3_fragility_gap_cases.csv` and `figure3_e4_pipeline_drift.csv`. The fragility gap, defined as evidence Spearman minus support Jaccard, is used solely for visualization.

Table F5: High-concordance fragile-support pairs. These pairs satisfy evidence Spearman ≥ 0.70 and support Jaccard < 0.20, showing that high evidence-rank concordance can coexist with unstable support regions.

Pair	Exp.	Source feature	Target feature	Target	EC	SO	DR	Gap	Overall
E4-028	E4	FS8 cortical volume	FastSurfer cortical volume	CN–MCI	0.896	0.026	0.188	0.871	fragile
E4-029	E4	FS8 cortical volume	FastSurfer cortical volume	MMSE	0.900	0.053	0.647	0.847	fragile
E4-027	E4	FS8 cortical volume	FastSurfer cortical volume	CN–AD	0.878	0.053	0.688	0.825	fragile
E4-032	E4	FS8 subcortical volume	FastSurfer subcortical volume	CN–MCI	0.948	0.143	0.529	0.805	fragile
E4-031	E4	FS8 subcortical volume	FastSurfer subcortical volume	CN–AD	0.926	0.143	0.563	0.783	fragile
E4-034	E4	FS8 subcortical volume	FastSurfer subcortical volume	age	0.940	0.176	0.353	0.764	fragile
E4-033	E4	FS8 subcortical volume	FastSurfer subcortical volume	MMSE	0.906	0.176	0.563	0.730	fragile

Table F6: FreeSurfer 8 versus FastSurfer row-level drift metrics. All E4 pairs are fragile, with low support overlap across cortical and subcortical volume comparisons despite generally high evidence concordance.

Target	Feature family	Source feature	Target feature	EC	SO	DR	Transfer	Transfer label	Overall
CN–AD	cortical volume	FS8 cortical volume	FastSurfer cortical volume	0.878	0.053	0.688	AUROC 0.564	weak	fragile
CN–MCI	cortical volume	FS8 cortical volume	FastSurfer cortical volume	0.896	0.026	0.188	AUROC 0.567	weak	fragile
MMSE	cortical volume	FS8 cortical volume	FastSurfer cortical volume	0.900	0.053	0.647	$\rho = -0.133$	weak	fragile
age	cortical volume	FS8 cortical volume	FastSurfer cortical volume	0.578	0.026	0.000	$\rho = -0.169$	weak	fragile
CN–AD	subcortical volume	FS8 subcortical volume	FastSurfer subcortical volume	0.926	0.143	0.563	AUROC 0.548	weak	fragile
CN–MCI	subcortical volume	FS8 subcortical volume	FastSurfer subcortical volume	0.948	0.143	0.529	AUROC 0.558	weak	fragile
MMSE	subcortical volume	FS8 subcortical volume	FastSurfer subcortical volume	0.906	0.176	0.563	$\rho = -0.113$	weak	fragile
age	subcortical volume	FS8 subcortical volume	FastSurfer subcortical volume	0.940	0.176	0.353	$\rho = -0.141$	weak	fragile

The cortical-volume age pair appears in Table F6. Its evidence Spearman is below the high-concordance threshold for inclusion in Table F5.

F.5 E4 robustness, uncertainty, and covariate sensitivity

The sensitivity and uncertainty analyses use the eight E4 FreeSurfer 8–FastSurfer claim-family comparisons (cortical and subcortical volume; CN–AD, CN–MCI, MMSE, and age). Both pipelines were evaluated on the same observations, target, and design matrix within each comparison. The primary support rule combines BH-FDR-supported regions with a top-k support fallback; statistical significance is assessed through BH-FDR. All results use exact ROI alignment, with 68 cortical or 64 subcortical matched ROIs and no ambiguous or duplicate matches. Literal-name and conservative one-to-one matching returned the same matched spaces.

Support and label sensitivity. Table F7 varies the top-k definition. Labels change for smaller support sets, whereas all eight comparisons are fragile at each tested $k = 20, 25, 30$. Under FDR-only support at $q \leq 0.05$, 0.10, and 0.20, none of the eight comparisons met moderate stability (some FDR supports were empty). Holding the continuous metrics fixed and varying only the reporting thresholds gave 3 stable and 5 fragile pairs under a permissive rule; the primary and more stringent rules each gave 8 fragile pairs. Categorical labels therefore depend on the declared thresholds.

Table F7: E4 top-k sensitivity under exact ROI matching (8 pairs per row).

k	**Stable**	**Moderate**	**Fragile**
5	1	4	3
10	0	3	5
15	0	2	6
20 (primary)	0	0	8
25	0	0	8
30	0	0	8

Null-overlap significance. We tested whether the source and target supports overlapped less than random sets of their observed sizes, using an exact lower-tail hypergeometric test in each matched ROI universe ($N = 68$ cortical; $N = 64$ subcortical). Benjamini–Hochberg correction was applied across all eight pair-level tests, separately from the ROI-wise BH-FDR procedure used in claim extraction. Table F8 reports the exact tests for the primary hybrid support rule at $k = 20$. All four cortical comparisons had significantly below-random support overlap ($q < 0.05$); no subcortical comparison did. The subcortical results are compatible with chance-level overlap and provide insufficient evidence for equivalence or support preservation.

Table F8: Exact lower-tail support-overlap null tests for E4 ($k = 20$); p is hypergeometric and q is BH-adjusted across eight comparisons.

Feature	**Target**	**Overlap**	**SO**	p	q
Cortical volume	CN–AD	2/20	0.053	0.0199	0.0399
Cortical volume	CN–MCI	1/20	0.026	0.0030	0.0121
Cortical volume	MMSE	2/20	0.053	0.0199	0.0399
Cortical volume	Age	1/20	0.026	0.0030	0.0121
Subcortical volume	CN–AD	5/20	0.143	0.3360	0.4480
Subcortical volume	CN–MCI	5/20	0.143	0.3360	0.4480
Subcortical volume	MMSE	6/20	0.176	0.5631	0.5631
Subcortical volume	Age	6/20	0.176	0.5631	0.5631

Note. Overlap is the number of ROIs shared by the two 20-region support sets; SO is their Jaccard overlap. Under this null, the expected intersection is $20 \times 20/N$ (5.88 for cortical and 6.25 for subcortical).

Paired bootstrap uncertainty. We used 300 paired bootstrap resamples per E4 comparison (2,400 valid resamples; zero failures), resampling identical observations for FreeSurfer 8 and FastSurfer. Table F9 reports percentile 95% confidence intervals for evidence Spearman, support Jaccard, and source-support direction retention. Evidence concordance had lower confidence limits of 0.760–0.912 in the seven higher-concordance comparisons; cortical age was weaker. Support membership at the

top-k boundary can change under resampling, giving wider or discrete overlap intervals. The exact matched-size null test evaluates below-random support overlap, while bootstrap intervals describe sampling variability.

Table F9: Paired bootstrap 95% percentile intervals for E4 (300 valid resamples per pair).

Feature	Target	EC 95% CI	SO 95% CI	DR 95% CI
Cortical volume	CN–AD	[0.813, 0.918]	[0.000, 0.600]	[0.118, 1.000]
Cortical volume	CN–MCI	[0.760, 0.908]	[0.000, 0.600]	[0.000, 1.000]
Cortical volume	MMSE	[0.809, 0.924]	[0.000, 0.667]	[0.111, 1.000]
Cortical volume	Age	[0.389, 0.714]	[0.000, 0.111]	[0.000, 0.000]
Subcortical volume	CN–AD	[0.841, 0.961]	[0.111, 0.481]	[0.342, 1.000]
Subcortical volume	CN–MCI	[0.813, 0.968]	[0.053, 0.635]	[0.188, 1.000]
Subcortical volume	MMSE	[0.801, 0.944]	[0.111, 0.538]	[0.312, 1.000]
Subcortical volume	Age	[0.912, 0.957]	[0.143, 0.212]	[0.256, 0.400]

Acquisition and site adjustment. Table 2 separates changes due to restricting the sample to acquisition-complete observations from changes due to adjusting for field strength, cohort, and categorical SITE effects. The last three specifications use identical row sets for each pair, with common targets and source/target design matrices. The median support Jaccard increased from 0.194 under the primary covariates on acquisition-complete observations to 0.510 after adding field strength and cohort, then changed to 0.404 in the categorical site models. The site designs included 58 or 61 SITE dummy columns (59 or 62 levels, respectively). The two CN–AD site models had rank 66 of 67 with condition numbers approximately 1.28×10^{23}; the remaining site designs were full rank but could be ill-conditioned (condition numbers on the order of 10^8). The site-effect models use pseudoinverse OLS. Rank deficiency in the CN–AD designs prevents identification of their individual site coefficients, while the high condition numbers in other designs require caution in interpreting those coefficients. The observed support profiles depend on the acquisition/covariate specification, which is recorded as part of each SCS analysis setting.

F.6 Atlas and feature-family sensitivity

Table F10 reports the representation-sensitivity values underlying Figure 4. The table aggregates E5 and E6 by comparison and target group. Rows with no evaluated pairs, including demographic/acquisition targets in E5/E6, are omitted. Weighted means use the number of evaluated pairs as weights. Missing EC and SO values indicate comparisons where the available feature spaces did not support meaningful evidence-concordance or support-overlap computation.

Table F10: Atlas and feature-family sensitivity summary. EC denotes weighted mean evidence Spearman and SO denotes weighted mean support Jaccard; stability counts are reported as stable/moderate/fragile/insufficient.

Exp.	Comparison	Target group	Pairs	Stability counts	EC	SO
E5	Kong17 → Yale696	diagnosis/impairment	5	2/2/1/0	0.680	1.000
E5	Kong17 → Yale696	age	3	0/0/3/0	0.371	1.000
E5	Kong17 → Yale696	cognition	3	1/1/1/0	0.600	1.000
E5	Schaefer7 → Kong17	diagnosis/impairment	5	5/0/0/0	0.966	1.000
E5	Schaefer7 → Kong17	age	3	2/1/0/0	0.771	1.000
E5	Schaefer7 → Kong17	cognition	3	3/0/0/0	0.943	1.000
E5	Schaefer7 → Yale696	diagnosis/impairment	5	2/0/3/0	0.520	1.000
E5	Schaefer7 → Yale696	age	3	0/2/1/0	0.600	1.000
E5	Schaefer7 → Yale696	cognition	3	1/1/1/0	0.500	1.000
E6	cortical volume → Schaefer7 thickness	diagnosis/impairment	5	2/1/2/0	0.520	1.000
E6	cortical volume → Schaefer7 thickness	age	3	0/0/3/0	0.300	1.000
E6	cortical volume → Schaefer7 thickness	cognition	3	2/0/1/0	0.700	1.000
E6	subcortical volume → Schaefer7 thickness	diagnosis/impairment	5	0/0/0/5	–	–
E6	subcortical volume → Schaefer7 thickness	age	3	0/0/0/3	–	–
E6	subcortical volume → Schaefer7 thickness	cognition	3	0/0/0/3	–	–

The support-overlap values in Table F10 measure preservation of coarse anatomical or network support under atlas or feature-family perturbation.

F.7 Anatomical claim summaries

These tables describe the anatomical distribution of supported claim cards generated by the claim-review pipeline. High counts identify anatomical categories with more supported statements; their interpretation as biological localization would require vertex-wise or region-wise evidence.

Table F11 summarizes supported claim counts by anatomical lobe and target group. For each lobe–target combination, the table reports both the number of supported claim cards and the median absolute evidence value, corresponding to the summary data used for Figure 4 Panels A–B. The largest number of supported claims appeared in temporal, frontal, parietal, and unknown/anatomically mixed regions, with temporal-lobe claims particularly frequent for diagnosis/impairment and cognition-related targets. These counts summarize the representation of broad anatomical categories among supported claims.

Table F11: Supported claim counts by lobe and target group. The table reports the claim-card counts and median absolute evidence values underlying Figure 4 Panels A–B.

Lobe	Target group	Claims	Median $\lvert e \rvert$
cingulate	age	27	4.299
cingulate	cognition	15	9.540
cingulate	demographic/acquisition	10	9.104
cingulate	diagnosis/impairment	36	6.749
frontal	age	153	4.202
frontal	cognition	25	4.216
frontal	demographic/acquisition	32	92.961
frontal	diagnosis/impairment	135	4.496
insula	age	38	3.503
insula	cognition	3	10.533
insula	demographic/acquisition	10	90.718
insula	diagnosis/impairment	23	6.709
medial	age	2	3.744
medial	diagnosis/impairment	4	5.191
occipital	age	41	5.961
occipital	cognition	15	8.268
occipital	demographic/acquisition	14	4.733
occipital	diagnosis/impairment	42	5.422
other_yale	age	6	3.309
other_yale	cognition	10	6.044
other_yale	diagnosis/impairment	12	5.903
parietal	age	37	4.162
parietal	cognition	93	7.933
parietal	demographic/acquisition	19	8.706
parietal	diagnosis/impairment	83	6.786
sensorimotor	age	18	3.393
sensorimotor	cognition	16	4.897
sensorimotor	diagnosis/impairment	17	4.420
temporal	age	120	6.040
temporal	cognition	194	9.203
temporal	demographic/acquisition	32	3.524
temporal	diagnosis/impairment	310	7.598
unknown	age	107	7.700
unknown	cognition	54	10.157
unknown	demographic/acquisition	49	9.199
unknown	diagnosis/impairment	159	6.787

Table F12 further summarizes supported cortical-thickness claim cards according to 7-network assignment, hemisphere, and target group. This table corresponds to the network–hemisphere summary used in Figure 4 Panel B. Across networks, Default, Control, Limbic, and Somatomotor systems contributed many supported cortical-thickness claims, although the distribution varied by hemisphere and target group. As in the lobe-level analysis, the counts describe the distribution of claim cards; network-specific vulnerability would require a direct statistical test.

Table F12: Cortical-thickness network and hemisphere claim counts. Counts are derived from supported cortical-thickness claim cards and summarize the network–hemisphere data used in Figure 4 Panel B.

Network	Hemisphere	Target group	Claims
Cont	LH	age	16
Cont	LH	cognition	25
Cont	LH	demographic/acquisition	2
Cont	LH	diagnosis/impairment	34
Cont	RH	age	27
Cont	RH	cognition	15
Cont	RH	demographic/acquisition	7
Cont	RH	diagnosis/impairment	31
Default	LH	age	31
Default	LH	cognition	59
Default	LH	demographic/acquisition	8
Default	LH	diagnosis/impairment	93
Default	RH	age	28
Default	RH	cognition	36
Default	RH	demographic/acquisition	8
Default	RH	diagnosis/impairment	61
DorsAttn	LH	age	6
DorsAttn	LH	cognition	15
DorsAttn	LH	demographic/acquisition	4
DorsAttn	LH	diagnosis/impairment	5
DorsAttn	RH	age	1
DorsAttn	RH	cognition	5
DorsAttn	RH	demographic/acquisition	4
DorsAttn	RH	diagnosis/impairment	6
Limbic	LH	age	12
Limbic	LH	cognition	15
Limbic	LH	demographic/acquisition	4
Limbic	LH	diagnosis/impairment	34
Limbic	RH	age	5
Limbic	RH	cognition	9
Limbic	RH	demographic/acquisition	3
Limbic	RH	diagnosis/impairment	25
SalVentAttn	LH	age	15
SalVentAttn	LH	cognition	5
SalVentAttn	LH	demographic/acquisition	6
SalVentAttn	LH	diagnosis/impairment	19
SalVentAttn	RH	age	19
SalVentAttn	RH	cognition	5
SalVentAttn	RH	demographic/acquisition	5
SalVentAttn	RH	diagnosis/impairment	19
SomMot	LH	age	33
SomMot	LH	demographic/acquisition	4
SomMot	LH	diagnosis/impairment	22
SomMot	RH	age	31
SomMot	RH	demographic/acquisition	2
SomMot	RH	diagnosis/impairment	24
Vis	LH	age	14
Vis	LH	cognition	6
Vis	LH	demographic/acquisition	14
Vis	LH	diagnosis/impairment	20
Vis	RH	age	17
Vis	RH	cognition	5
Vis	RH	demographic/acquisition	17
Vis	RH	diagnosis/impairment	17

F.8 Expert and LLM review outcomes

Tables F13–F16 provide the numerical review summaries underlying Figure 4 Panels D–F. The tables report final human expert decisions and reviewer scores for claim plausibility, clarity, faithfulness, usefulness, and reviewer confidence. Rows with `insufficient_reviews` identify LLM-only extended claims without complete human review.

Table F13 reports the final decision distribution for the 60 claims with complete human expert review. Most reviewed claims were judged to require either major or minor revision, while a smaller subset was accepted without revision or rejected. The review indicates that generated claims were often scientifically useful and required expert correction, refinement, or qualification before publication.

Table F13: Human expert decision counts. Counts include only claims with complete human expert review.

Expert decision	Claims
expert_major_revision	19
expert_minor_revision	17
expert_accepted	15
expert_rejected	5
expert_disagreement	4
Total human-reviewed claims	**60**

Table F14 summarizes the mean 1–5 scores assigned by each reviewer across the five review dimensions. The LLM reviewers tended to assign high clarity scores, whereas human experts showed more variation across domains and generally provided more conservative assessments. These results support auxiliary LLM review for large-scale screening, with human expert review as the reference for scientific judgment.

Table F14: Mean review scores by reviewer and dimension. Scores use a 1–5 scale, with LLM reviewers providing auxiliary assessments alongside human experts.

Reviewer	Type	Plaus.	Clarity	Faith.	Useful.	Conf.
chatgpt_5_5_thinking	LLM	3.875	4.783	4.525	3.625	4.175
gemini_3_1_pro	LLM	4.225	4.950	3.942	4.408	3.892
geriatrics	human	4.150	3.617	3.750	3.967	3.517
neuroscience	human	3.683	2.883	3.550	3.500	2.767
radiology	human	4.300	3.467	4.100	4.117	4.533

Table F15 reports agreement among the three human experts. Agreement was highest for plausibility and clarity, moderate for faithfulness and usefulness, and lowest for reviewer confidence. The comparison of ICC$(2,1)$ with ICC$(2,k)$ indicates higher reliability for average scores across human reviewers than for individual scores.

Table F15: Human agreement by review dimension. ICC$(2,1)$ estimates single-reviewer reliability, while ICC$(2,k)$ estimates reliability of the average human-reviewer score.

Dimension	Claims	Reviewers	ICC$(2,1)$	ICC$(2,k)$	Pearson	Spearman
plausibility_1_5	60	3	0.426	0.690	0.475	0.400
clarity_1_5	60	3	0.398	0.665	0.525	0.506
faithfulness_1_5	60	3	0.360	0.628	0.379	0.393
usefulness_1_5	60	3	0.291	0.552	0.292	0.254
reviewer_confidence_1_5	60	3	0.106	0.263	0.296	0.295

Table F16 compares each LLM reviewer with the mean human expert score on the 60 fully reviewed claims. The strongest LLM–human rank agreement was observed for ChatGPT on plausibility, while agreement was weaker for clarity and several Gemini dimensions. Both LLM reviewers showed positive bias for clarity, rating claim wording more favorably than human experts.

Table F16: Human–LLM agreement by review dimension. Agreement is measured between each LLM reviewer and the mean human expert score on the 60 claims with complete human review.

LLM reviewer	Dimension	Overlap	Spearman	MAD	Bias
chatgpt_5_5_thinking	plausibility_1_5	60	0.670	0.594	-0.128
chatgpt_5_5_thinking	clarity_1_5	60	0.292	1.378	1.278
chatgpt_5_5_thinking	faithfulness_1_5	60	0.429	0.889	0.700
chatgpt_5_5_thinking	usefulness_1_5	60	0.519	0.778	-0.178
chatgpt_5_5_thinking	reviewer_confidence_1_5	60	0.322	0.733	0.611
gemini_3_1_pro	plausibility_1_5	60	0.439	0.706	0.006
gemini_3_1_pro	clarity_1_5	60	0.161	1.589	1.578
gemini_3_1_pro	faithfulness_1_5	60	0.163	1.061	0.317
gemini_3_1_pro	usefulness_1_5	60	0.163	0.928	0.572
gemini_3_1_pro	reviewer_confidence_1_5	60	0.351	0.689	0.278