Beyond Dice: Clinically Structured Bias Analysis of Multiple Sclerosis Lesion Segmentation Models

Abdul Basit ABDUL.BASIT@NYU.EDU
eBRAIN Lab, Division of Engineering, New York University (NYU) Abu Dhabi
Abu Dhabi, United Arab Emirates

Namik Hassan MM13924@NYU.EDU
eBRAIN Lab, Division of Engineering, New York University (NYU) Abu Dhabi
Abu Dhabi, United Arab Emirates

Muhammad Shafique MUHAMMAD.SHAFIQUE@NYU.EDU
eBRAIN Lab, Division of Engineering, New York University (NYU) Abu Dhabi
Abu Dhabi, United Arab Emirates

Abstract

Automated multiple sclerosis (MS) lesion segmentation is commonly benchmarked with voxel-overlap metrics such as Dice, although clinical interpretation also depends on lesion size, anatomical location, and the preservation of subject-level dissemination evidence. We present *Beyond-Dice*, a reproducible MS-specific evaluation framework spanning voxel fidelity, lesion detection, anatomical location, subject-level evidence, case-type stress testing, and split/merge topology. We compare nnU-Net, SegResNet, and SwinUNETR on a separate 24-subject MSSEG-1 evaluation cohort containing 1,172 lesions. Subject-bootstrap 95% confidence intervals (CIs) show statistically tied Dice: 0.713 [0.648, 0.772] for nnU-Net, 0.713 [0.669, 0.759] for SegResNet, and 0.708 [0.656, 0.756] for SwinUNETR. Non-overlap endpoints nevertheless separate clinically relevant behavior: nnU-Net improves subject-mean lesion recall over SegResNet by 0.051 [0.016, 0.088], whereas SwinUNETR produces 10.67 additional false-positive lesions per scan relative to nnU-Net [5.88, 16.38]. Across all models, 238 lesions are missed by all three and 226 have model-dependent detection; lesion volume is the dominant detection predictor (odds ratio 3.86 per log-volume unit [3.45, 4.74]). A modality-matched MSLesSeg-to-MSSEG-1 experiment further shows similar matched-lesion Dice across locations (0.584–0.659) despite markedly lower infratentorial recall/F1 (0.339/0.412). These results operationalize how Dice-tied models can preserve different clinical evidence and support endpoint-specific, uncertainty-aware model selection rather than a single-score leaderboard.

1. Introduction

Multiple sclerosis (MS) is a chronic inflammatory and neurodegenerative disease in which MRI-visible lesions are central to diagnosis, staging, and longitudinal monitoring. Clinical interpretation depends not only on lesion volume but also on lesion topography. Periventricular (PV), juxtacortical/cortical (JC), and infratentorial (IT) findings contribute differently to dissemination-in-space (DIS) reasoning and clinical management (Wattjes et al., 2015; Ghasemi et al., 2017; Haki et al., 2024). A segmentation model can therefore achieve strong voxel overlap while failing to preserve the evidence needed for a particular downstream use.

1.1. State of the Art and Evaluation Gap

Automated MS lesion segmentation has advanced from early challenge benchmarks and classical approaches to modern deep networks (Styner et al., 2008; García-Lorenzo et al., 2013; Lladó et al., 2012; Danelakis et al., 2018; Zeng et al., 2020). Architectures such as nnU-Net, SegResNet, and SwinUNETR provide strong aggregate benchmark performance (Isensee et al., 2018; Myronenko, 2018; Hatamizadeh et al., 2022). Global Dice and HD95 remain necessary summaries, but they do not identify which lesions are missed, whether errors concentrate in a clinically relevant location, whether a predicted topographic pattern is preserved at subject level, or whether one-to-one matching is penalizing a split/merge representation rather than a complete miss. This is a problem of endpoint specification, consistent with broader recommendations that validation metrics be selected from the clinical task and target structure rather than by convention alone (Maier-Hein et al., 2024).

The resulting translation gap has three practical sources. First, overlap metrics are volume dominated, so errors on small components may have little influence on a scan-level score. Second, global overlap is topography agnostic. Third, conventional one-to-one lesion matching can conflate detection failure with split/merge topology. These mechanisms can make models with similar Dice behave differently for lesion detection, topographic evidence preservation, and false-positive burden.

1.2. Problem Formulation

We ask a deliberately scoped question: *when MS lesion segmenters have similar global Dice, what additional behavior becomes visible when evaluation follows the clinical evidence structure?* We operationalize this question through five endpoint families: lesion size, anatomical location, subject-level topographic evidence, clinically sensitive case types, and split/merge topology. The PV/JC/IT ontology and DIS proxy are MS-specific. The transferable contribution is the evaluation recipe: define task-relevant evidence units, map predictions to those units, quantify uncertainty at the correct sampling level, and report trade-offs rather than infer deployment suitability from one aggregate score.

Our controlled comparison uses three model families on a separate 24-subject MSSEG-1 evaluation cohort. We additionally test endpoint persistence with MSLesSeg, an independent MS dataset containing 93 scans from 53 patients. The external experiment is modality matched: an nnU-Net trained on MSLesSeg T1/T2/FLAIR is evaluated on MSSEG-1 using those same three modalities, rather than the five modalities used by the primary MSSEG-1 models. This is an additional MS-domain stress test, not a cross-disease claim or an external re-ranking of all architectures.

1.3. Generalizable Insights about Machine Learning in the Context of Healthcare

This study yields four insights that extend beyond MS while preserving the clinical scope of the evidence. First, the *unit of evaluation should match the unit of decision*: a voxel score cannot by itself validate lesion detection or subject-level evidence preservation. Second, healthcare ML models should be compared as endpoint-specific trade-offs; a model that is preferable for sensitivity-oriented screening may be inappropriate for a precision-oriented

monitoring workflow. Here, "screening" denotes a use that prioritizes finding possible lesion evidence for subsequent expert review, whereas "monitoring" denotes repeated measurement where false additions can distort apparent change. Neither use is evaluated as a clinical trial in this work. Third, uncertainty must respect data hierarchy: subjects are the independent sampling units, while lesions are clustered within subjects. Fourth, rare clinical presentations are valuable safety stress tests but cannot support population-level comparative conclusions when subgroup counts are very small.

The generalizable object is therefore not the MS-specific DIS rule. It is an auditable evaluation skeleton that separates voxel fidelity, object detection, anatomical-context preservation, subject-level evidence, and structural topology. Applying it to another disease requires a new clinically justified ontology and decision rule; it is not a plug-and-play claim of cross-disease validity.

1.4. Novel Contributions

To address this gap, we propose *Beyond-Dice*, an MS-specific evidence-preservation evaluation framework, with the following contributions:

- We introduce an MS-specific, architecture-agnostic evidence-preservation benchmark that unifies voxel, lesion, location, subject, case-type, and topology endpoints.
- We provide an auditable anatomical labeling pipeline with explicit FreeSurfer labels, physical-space dilation, secondary labels, overlap/distance quality-control fields, and sensitivity analyses using SAMSEG and minimum-contact thresholds.
- We quantify uncertainty with paired subject bootstrap, subject-cluster lesion bootstrap, exact binomial intervals, and leave-one-subject-out rank stability.
- We empirically trace the mechanism hidden by tied Dice using lesion-survival analysis and within-subject detection patterns across 1,172 lesions.
- We add same-domain and cross-dataset MSLesSeg validation, showing that lesion-size and location-specific endpoint gaps persist under an independent MS dataset and a modality-matched dataset shift.

The goal is not to replace established segmentation metrics. It is to make their clinical meaning explicit by reporting what evidence is preserved, where it is lost, and how model preference changes with the intended endpoint.

2. Background and Related Work

2.1. From Segmentation Accuracy to Clinical Evidence Preservation

MS lesion segmentation has evolved from classical intensity/atlas pipelines to deep-learning systems with strong benchmark performance (García-Lorenzo et al., 2013; Zeng et al., 2020; Brosch et al., 2016; Valverde et al., 2017; Aslani et al., 2019). Recent work emphasizes robustness across scanners and cohorts (Hindsholm et al., 2023; van Nederpelt et al., 2024; Badea et al., 2025) and challenge-driven comparisons (Rondinella et al., 2024; Guarnera et al., 2025). Yet, as also highlighted in prior surveys (Danelakis et al., 2018; Kaur et al.,

2020; Ma et al., 2022; Nicolaou et al., 2024), most evaluations still center on overlap-oriented endpoints (Dice/HD95, lesion detection F1), which are necessary but not sufficient for clinical trust.

This motivation is aligned with Metrics Reloaded, which establishes that metric selection should be problem-aware and reflect the domain interest, target structure, and output type (Maier-Hein et al., 2024). Our contribution is complementary and MS-specific: we instantiate that principle as an auditable endpoint decomposition, then empirically trace how Dice-tied models differ in lesion survival, anatomical evidence, subject-level DIS-proxy preservation, false-positive burden, and split/merge topology.

For this paper, the critical question is not architectural novelty, but *evaluation completeness*: does a method quantify clinically relevant evidence preservation beyond aggregate overlap? To make this explicit, Table 1 compares representative and influential studies against *both* conventional endpoints (where many prior works are strong) and clinically structured endpoints (where gaps remain).

Table 1: Comparison of representative MS lesion segmentation studies. Columns indicate whether each work reports the corresponding evaluation capability (✓: covered, ✗: not explicitly covered).

Representative work	Voxel overlap	Lesion-wise det.	Boundary/ vol.	Size-aware	Location-aware	Subject-level	Case-type stratified	Split/ merge-aware	Rank shifts
Johnston et al. (1996)	✓	✗	✗	✗	✗	✗	✗	✗	✗
Egger et al. (2017)	✓	✗	✓	✓	✗	✗	✗	✗	✓
Jain et al. (2016)	✓	✓	✓	✓	✗	✗	✗	✗	✓
Carass et al. (2017)	✓	✓	✓	✗	✗	✓	✗	✗	✓
Commowick et al. (2018)	✓	✓	✓	✓	✗	✗	✗	✓	✓
Aslani et al. (2019)	✓	✓	✓	✗	✗	✗	✗	✗	✓
McKinley et al. (2019)	✓	✓	✓	✗	✗	✗	✗	✓	✓
McKinley et al. (2021)	✓	✓	✗	✗	✗	✗	✗	✓	✓
Zhang et al. (2021)	✓	✓	✓	✗	✗	✗	✗	✗	✓
Hitziger et al. (2022)	✓	✓	✗	✗	✗	✗	✗	✗	✓
Hindsholm et al. (2023)	✓	✓	✗	✗	✗	✗	✗	✗	✓
Wu et al. (2023)	✓	✓	✓	✗	✗	✗	✗	✗	✓
Rosa et al. (2024)	✓	✓	✓	✓	✗	✗	✗	✗	✓
van Nederpelt et al. (2024)	✓	✓	✓	✗	✓	✗	✗	✗	✓
Badea et al. (2025)	✓	✗	✗	✗	✗	✗	✗	✗	✗
Rondinella et al. (2024)	✓	✗	✗	✗	✗	✗	✗	✗	✗
Guarnera et al. (2025)	✓	✓	✓	✗	✗	✗	✗	✗	✓
This work	✓	✓	✓	✓	✓	✓	✓	✓	✓

2.2. What the Comparison Reveals

Table 1 highlights a recurring pattern. Prior work is strong on *conventional evaluation* (voxel overlap, lesion-wise detection, and in several cases boundary/volume reporting). Several studies report different "best" methods depending on whether the endpoint is overlap, lesion detection, or volumetric/boundary performance. However, these strengths still do not generally extend to *clinically structured evaluation*: anatomical location-aware reporting, subject-level dissemination evidence preservation, case-type stratified failure analysis, and explicit split/merge topology analysis remain sparse. This leaves an important blind spot: models can appear similar on aggregate metrics while differing in clinically consequential failure modes.

Three gaps are especially persistent. First, **clinical-topography reporting is sparse**: PV/JC/IT behavior is seldom quantified in a unified evaluation framework, despite its diagnostic importance (Wattjes et al., 2015). Second, **subject-level evidence analysis**

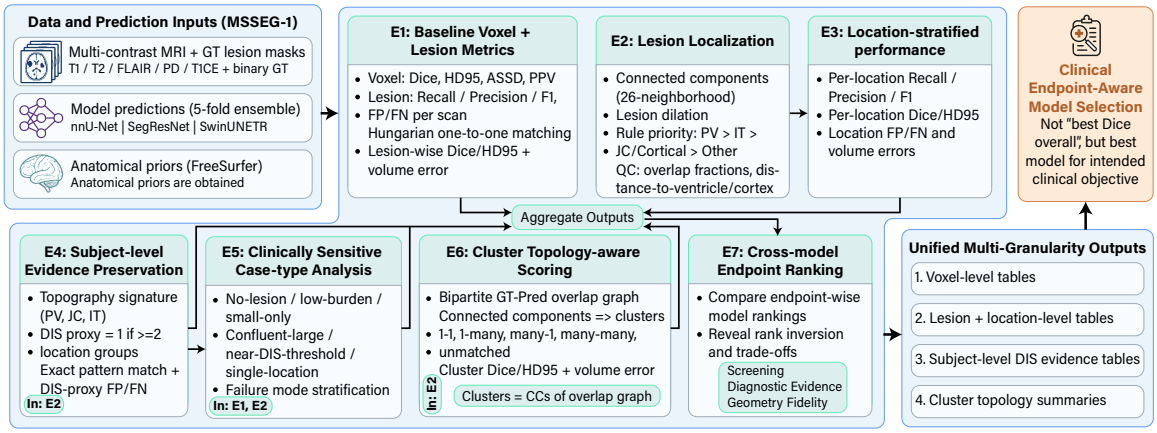

Figure 1: Overview of the proposed clinically structured E1–E7 evaluation pipeline. Inputs (GT masks, model predictions, and FreeSurfer priors) are transformed into voxel-, lesion-, location-, subject-, and cluster-level endpoints, and finally consolidated for endpoint-aware cross-model ranking.

is rare: most studies stop at voxel/lesion aggregates without measuring preservation of dissemination patterns. Third, **error-structure disambiguation remains limited**: only a minority explicitly encode split/merge-aware matching logic in their evaluation design (Fartaria et al., 2018). This unified design directly tests the central hypothesis of this paper: *model ranking in MS lesion segmentation is endpoint-dependent, and overlap-only ranking can be clinically misleading.*

3. Methodology

3.1. Pipeline Overview

We cast MS lesion segmentation evaluation as a **clinical evidence preservation** problem rather than a single overlap-optimization problem. Let a model produce a binary lesion prediction mask for each subject. Our framework evaluates that prediction across four coupled levels: voxel fidelity, lesion detection/geometry, subject-level topographic evidence, and structural topology under split/merge behavior.

The framework is organized as seven experiments (E1–E7): baseline voxel/lesion metrics (E1), anatomically guided lesion localization (E2), location-specific bias analysis (E3), subject-level dissemination-evidence preservation (E4), clinically sensitive case-type stress testing (E5), topology-aware cluster evaluation (E6), and endpoint-wise multi-model ranking (E7). Statistical uncertainty, label robustness, lesion-survival modeling, and external MS validation are applied across these experiments. As summarized in Figure 1, this decomposition separates *where* errors occur, *how* they alter subject-level evidence, and *whether* model preference depends on the deployment endpoint.

3.2. Data Representation and Lesion Units

For each subject s, lesions are represented as 3D connected components (26-neighborhood). Let $G^{(s)} = \{g_i\}_{i=1}^{N_G}$ and $P^{(s)} = \{p_j\}_{j=1}^{N_P}$ denote GT and predicted lesion sets. We define the overlap matrix

$$\mathbf{O}^{(s)} \in \mathbb{N}^{N_G \times N_P}, \qquad O_{ij}^{(s)} = |g_i \cap p_j|, \tag{1}$$

which induces all lesion-level correspondences used in E1 and E6.

All analysis masks are represented on an isotropic $1\,\mathrm{mm}^3$ grid, so voxel count equals physical volume in mm^3. Lesions are grouped as **small**: $v < 10\,\mathrm{mm}^3$, **medium**: $10 \leq v < 100\,\mathrm{mm}^3$, and **large**: $v \geq 100\,\mathrm{mm}^3$. These are operational analysis bins, not diagnostic thresholds. They expose size-dependent metric behavior and are reported with their denominators (194/708/270 lesions in MSSEG-1). The need to report size explicitly is supported by prior evidence that small lesions form a substantial component of MS lesion distributions and are vulnerable to volume-dominated summaries (Wang et al., 1997).

3.3. Experiment 1: Baseline Voxel and Lesion Metrics

E1 establishes a conventional but necessary baseline. At voxel level, we compute Dice, HD95, ASSD, and PPV. At lesion level, detection is overlap-based:

$$\mathrm{TP}_\ell = \#\{g_i : \exists j,\ O_{ij} > 0\}, \quad \mathrm{FN}_\ell = N_G - \mathrm{TP}_\ell, \quad \mathrm{FP}_\ell = \#\{p_j : \forall i,\ O_{ij} = 0\}, \tag{2}$$

from which lesion recall, precision, F1, and FP/FN per scan are derived.

To characterize geometric fidelity after detection, we solve one-to-one lesion assignment with the Hungarian method (Kuhn, 1955) over overlap-supported pairs (cost $= 1 - \mathrm{Dice}$). Matched pairs are then evaluated with lesion-wise Dice/HD95 and absolute/relative volume error. This separation avoids conflating detection failures with boundary-quality failures. The matching procedure is summarized in Algorithm 1.

Algorithm 1: One-to-One Lesion Matching (Hungarian) in E1

Input: GT components G, predicted components P, overlap matrix $\mathbf{O}$
Output: Matched index set $\mathcal{M} \subseteq \{1,\dots,N_G\} \times \{1,\dots,N_P\}$

$N_G \leftarrow |G|$, $N_P \leftarrow |P|$ Initialize $\mathbf{C} \in \mathbb{R}^{N_G \times N_P}$ with $+\infty$ **for** $i \leftarrow 1$ **to** N_G **do**
  $v_i \leftarrow |g_i|$
for $j \leftarrow 1$ **to** N_P **do**
  $u_j \leftarrow |p_j|$
for $i \leftarrow 1$ **to** N_G **do**
  for $j \leftarrow 1$ **to** N_P **do**
    if $O_{ij} > 0$ ***and*** $(v_i + u_j) > 0$ **then**
      $D_{ij} \leftarrow \dfrac{2O_{ij}}{v_i + u_j}$ $C_{ij} \leftarrow 1 - D_{ij}$
$\mathcal{A} \leftarrow$ HUNGARIAN($\mathbf{C}$) $\mathcal{M} \leftarrow \emptyset$ **foreach** $(i,j) \in \mathcal{A}$ **do**
  if $O_{ij} > 0$ **then**
    $\mathcal{M} \leftarrow \mathcal{M} \cup \{(i,j)\}$
return $\mathcal{M}$

3.4. Experiment 2–3: Anatomically Guided Lesion Location Labeling and Location Bias

E2 maps each lesion to one primary anatomical class among {PV, JC, IT, other}. Subject-specific anatomy is generated with FreeSurfer 8.1.0 (Fischl, 2012). From `aseg.mgz`, ventricle

labels are {4,5,14,15,43,44,72} and infratentorial labels are {7,8,16,46,47}; cortex is defined by nonzero `ribbon.mgz`. Label maps are resampled to the lesion grid with each image's geometry/affine and nearest-neighbor interpolation. No new affine or deformable registration is estimated. A 3 mm physical-space cube dilation is applied before overlap tests to reduce brittleness at boundary voxels, followed by a deterministic contact-priority rule:

`periventricular` > `infratentorial` > `juxtacortical/cortical` > `other`.

Auxiliary attributes (distance to ventricle/cortex, overlap fractions, multiregion flag, and secondary label) are retained for quality control (QC) and sensitivity analyses. "Exact lesion-label agreement" means that the same connected component receives the same primary class from two anatomical engines. We assess robustness in two ways: (i) repeat labeling with sequence-adaptive SAMSEG priors (Puonti et al., 2016); and (ii) require minimum contact fractions of 0.5%, 1%, 2%, or 5% before applying the same priority order. The latter tests whether conclusions depend on a single-voxel contact without tuning a threshold to maximize agreement. The complete primary rule is given in Algorithm 2.

Algorithm 2: Rule-Based Lesion Location Labeling (FreeSurfer-guided)

Input: Lesion mask ℓ, spacing s, ventricle mask V, cortical mask C, infratentorial mask I
Output: Primary label, secondary label, QC attributes
Dilate lesion by 3 mm cube footprint in physical space Compute overlaps with V, C, I and overlap fractions **if** *overlap with* $V > 0$ **then**
  primary ← periventricular
else
  if *overlap with* $I > 0$ **then**
    primary ← infratentorial
  else
    if *overlap with* $C > 0$ **then**
      primary ← juxtacortical/cortical
    else
      primary ← other
Set secondary label to highest non-primary non-zero overlap class Compute distance-to-ventricle and distance-to-cortex (mm) **return** labels + QC fields

E3 then evaluates performance *conditioned on location class*. It reports recall, precision, F1, FP/FN counts, Dice, HD95, and volume errors per class, turning aggregate segmentation quality into clinically interpretable topographic behavior.

3.5. Experiment 4: Subject-Level Evidence Preservation

E4 shifts from lesion-level correctness to patient-level evidence preservation. For each subject, we construct a topographic signature over clinically relevant classes (PV/JC/IT). Let $T_{GT}, T_{Pred} \in \{0, 1, 2, 3\}$ denote how many of these three location groups are present in GT and prediction, respectively. Concretely, define binary presence indicators for each class $k \in \{\text{PV}, \text{JC}, \text{IT}\}$:

$$x_k^{GT} = \mathbf{1}_{\{\text{at least one GT lesion in class } k\}}, \qquad x_k^{Pred} = \mathbf{1}_{\{\text{at least one predicted lesion in class } k\}}. \quad (3)$$

Then $T_{GT} = x_{PV}^{GT} + x_{JC}^{GT} + x_{IT}^{GT}$, $\quad T_{Pred} = x_{PV}^{Pred} + x_{JC}^{Pred} + x_{IT}^{Pred}$. Thus, $T = 0$ means no clinically relevant class is present, $T = 1$ means one class is present, and $T \in \{2, 3\}$ means dissemination across at least two classes. We define a DIS-like binary proxy: DIS-proxy $= \mathbf{1}_{\{T \geq 2\}}$, which equals 1 when at least two clinically relevant location groups are present and

0 otherwise. In practice, this is computed separately for GT and prediction (i.e., using T_{GT} and T_{Pred}) and compared at subject level. For example, if GT lesions appear in PV and IT, then $T_{GT} = 2$ and DIS-proxy$(GT) = 1$; if prediction contains only PV lesions, then $T_{Pred} = 1$ and DIS-proxy$(Pred) = 0$ (a subject-level FN for dissemination evidence).

We report exact topography-count agreement, exact location-pattern agreement, DIS-proxy agreement, and directional DIS errors (FP/FN). These endpoints quantify whether a model preserves subject-level clinical evidence, not just local overlap.

3.6. Experiment 5: Hard-Case Stratification

E5 creates descriptive audit strata from GT characteristics:

- No lesion scans,
- Very low burden (default: 1–3 lesions),
- Only small lesions,
- Large confluent lesions (largest-lesion fraction threshold),
- Near DIS threshold ($T_{GT} = 2$),
- Only one clinically important location ($T_{GT} = 1$).

For each case type, we record detection, geometric fidelity, false positives, and subject-level evidence metrics. These rows are safety-oriented stress tests, not prevalence estimates. In particular, groups with $n = 2$ or $n = 5$ are not used for model-level comparative conclusions or statistical inference; their role is limited to identifying candidate failure examples for audit.

3.7. Experiment 6: Cluster-Based Topology-Aware Evaluation

E6 addresses a known limitation of one-to-one lesion matching: split/merge configurations are structurally many-to-many. We therefore define a bipartite overlap graph with GT and predicted lesions as nodes and non-zero overlaps as edges. Connected components of this graph form topology-aware lesion *clusters*. For each cluster, we compute:

- Cluster type (one-to-one, one-to-many, many-to-one, many-to-many, GT-only, Pred-only),
- Aggregated cluster Dice (union GT vs union prediction),
- Average cluster HD95,
- Absolute and relative cluster volume errors,
- Cluster location label (from GT union, or prediction union if GT-only absent).

Let $U_G(c)$ and $U_P(c)$ denote voxel unions of GT and prediction inside cluster c. Cluster Dice is: $\text{Dice}_{\text{cluster}}(c) = \frac{2|U_G(c) \cap U_P(c)|}{|U_G(c)| + |U_P(c)|}$, which measures structural agreement after resolving split/merge groupings into unified components. Cluster extraction from the bipartite overlap graph is detailed in Algorithm 3.

Algorithm 3: Cluster Construction from GT–Prediction Overlap Graph

Input: Overlap matrix $\mathbf{O}$ between GT lesions and predicted lesions
Output: Cluster set $\mathcal{C} = \{c_q\}_{q=1}^{Q}$ with node partitions and type labels

$N_G, N_P \leftarrow \dim(\mathbf{O})$ Define bipartite graph $\mathcal{G} = (V_G \cup V_P, E)$ where $V_G = \{g_1, \ldots, g_{N_G}\}$ and $V_P = \{p_1, \ldots, p_{N_P}\}$
$E \leftarrow \emptyset$ **for** $i \leftarrow 1$ **to** N_G **do**
  for $j \leftarrow 1$ **to** N_P **do**
    if $O_{ij} > 0$ **then**
      $E \leftarrow E \cup \{(g_i, p_j)\}$
$\{\mathcal{K}_q\}_{q=1}^{Q} \leftarrow$ CONNECTEDCOMPONENTS($\mathcal{G}$) $\mathcal{C} \leftarrow \emptyset$ **for** $q \leftarrow 1$ **to** Q **do**
  $G_q \leftarrow \mathcal{K}_q \cap V_G$, $P_q \leftarrow \mathcal{K}_q \cap V_P$ $n_G \leftarrow |G_q|$, $n_P \leftarrow |P_q|$ **if** $n_G = 1$ ***and*** $n_P = 1$ **then**
    $\tau_q \leftarrow$ one-to-one
  else if $n_G = 1$ ***and*** $n_P > 1$ **then**
    $\tau_q \leftarrow$ one-to-many
  else if $n_G > 1$ ***and*** $n_P = 1$ **then**
    $\tau_q \leftarrow$ many-to-one
  else if $n_G > 1$ ***and*** $n_P > 1$ **then**
    $\tau_q \leftarrow$ many-to-many
  else if $n_G > 0$ ***and*** $n_P = 0$ **then**
    $\tau_q \leftarrow$ GT-only
  else
    $\tau_q \leftarrow$ Pred-only
  $\mathcal{C} \leftarrow \mathcal{C} \cup \{(G_q, P_q, \tau_q)\}$
return $\mathcal{C}$

3.8. Experiment 7: Multi-Model Clinical Comparison and Ranking

E7 integrates outputs from E1/E3/E4/E5/E6 across models into a unified endpoint matrix spanning voxel fidelity, lesion detection, topography-aware sensitivity, subject-level dissemination evidence, and cluster topology quality. Rankings are computed per endpoint family to test whether model preference is stable or endpoint-dependent. We do not interpret negligible point-estimate differences as evidence of a universal ordering; the stable target is whether one model is Pareto-dominant across endpoint families.

3.9. Statistical Analysis and Mechanism Tracing

Subjects are the independent sampling units. For a subject-level metric m_s on $n = 24$ held-out subjects, we draw $B = 5{,}000$ bootstrap samples of subjects with replacement, compute:

$$\bar{m}^{(b)} = \frac{1}{n}\sum_{i=1}^{n} m_{s_i^{(b)}}, \qquad \mathrm{CI}_{95\%} = \left[P_{2.5}(\bar{m}^{(b)}), P_{97.5}(\bar{m}^{(b)})\right], \tag{4}$$

and report percentile confidence intervals (CIs) (Efron and Tibshirani, 1993). Model differences are bootstrapped on paired subjects, preserving the within-subject comparison. Lesion-level and location-level intervals use a subject-cluster bootstrap: a sampled subject contributes all of its lesions, so within-subject lesion dependence is retained. Binary subject endpoints use exact Clopper–Pearson intervals (Clopper and Pearson, 1934). We also repeat the endpoint ranking after leaving out each subject in turn.

To identify what is hidden by tied Dice, we align the 1,172 GT components across models and record eight detection patterns (detected/missed by each architecture). A pooled logistic lesion-detection model includes log lesion volume, distance to ventricle, distance to cortex, subject lesion burden, location, model indicators, and burden-by-location interactions. Uncertainty is clustered by subject. This analysis is explanatory and cohort-conditioned; it does not establish a universal biological detection law.

Aggregation convention: Unless a table is explicitly labeled "pooled lesions," metrics are first computed per subject and then averaged. Pooled lesion analyses report raw GT denominators and use subject-cluster uncertainty. All models share the same subjects, preprocessing, and endpoint definitions.

4. Experimental Setup

4.1. MSSEG-1 Cohort and Split

We use the full-access MICCAI 2016 MS lesion segmentation benchmark obtained through the official data portal and agreement. The controlled benchmark spans three centers, four scanners, 1.5T/3T acquisitions, five co-registered MRI contrasts (T1, T2, FLAIR, PD, and T1CE), and consensus lesion annotations derived from seven experts (Commowick et al., 2018). Model development uses 23 labeled cases under a shared five-fold split. A separate 24-subject cohort is used once for the reported comparison. The evaluation cohort contains 1,172 GT lesions: 194 small, 708 medium, and 270 large; anatomically, 189 are PV, 924 JC/cortical, and 59 IT. IT lesions occur in 12 of 24 subjects. These denominators support failure-mechanism analysis but do not turn the 24 subjects into a population-level leaderboard.

Inputs are resampled to isotropic 1-mm voxels using the nnU-Net preprocessing geometry. Intensities are normalized per channel over nonzero voxels. The MSSEG-1 modalities are supplied co-registered. Binary predictions are generated on the common lesion grid and evaluated against the same consensus masks.

4.2. Compared Models and Training Protocol

We benchmark nnU-Net (Isensee et al., 2018), SegResNet (Myronenko, 2018), and SwinUNETR (Hatamizadeh et al., 2022), representing self-configuring U-Net, residual convolutional, and transformer-hybrid families. All use the same five folds, 1,000-epoch budget, mixed-precision training, and no hyperparameter sweep. The default `nnUNetTrainer` is retained as the reproducible nnU-Net baseline; no alternative-trainer ablation is claimed. The complete losses, optimizers, schedules, crop sizes, regularization, architecture settings, and inference procedure are reported in Appendix Table 7.

SegResNet and SwinUNETR use positive/negative label-guided crops, independent flips along three axes ($p = 0.5$ each), random 90-degree rotations ($p = 0.5$), and intensity scale/shift perturbations of 0.1 ($p = 0.5$). Their random seed is 42, validation is performed every 10 epochs, and the final fold predictions are probability averaged before argmax. nnU-Net uses its planned augmentation and native folds 0–4 prediction command. Training ran on three NVIDIA RTX 5880 Ada GPUs in a common software environment.

4.3. Anatomical Processing and Postprocessing

FreeSurfer 8.1.0 `recon-all -all -parallel -openmp` is run on each T1 image; fresh cases additionally use `-i <T1>`. Experiments consume `aseg.mgz` and `ribbon.mgz`. Discrete labels are brought to lesion-mask space with geometry/affine-based nearest-neighbor resampling (interpolation order 0), not an estimated affine or deformable registration. Every lesion

record stores anatomical overlap fractions, ventricle/cortex distances, a multiregion flag, and a secondary label.

The primary results apply no small-component removal, lesion filtering, anatomical mask, or other cleanup. SegResNet and SwinUNETR use softmax fold ensembling followed by argmax; nnU-Net uses native `nnUNetv2_predict`. Component removal is evaluated only as a sensitivity analysis and is never folded into the reported primary masks.

4.4. Independent MS-Domain Validation

MSLesSeg contains 93 scans from 53 patients with T1, T2, and FLAIR (Guarnera et al., 2025). Fold 0 uses 74 scans for training and 19 for validation. We report two nnU-Net analyses. First, same-domain fold-0 validation tests whether size-dependent endpoints appear within MSLesSeg. Second, a 50-epoch MSLesSeg fold-0 model is applied to the 24-subject MSSEG-1 cohort using only T1/T2/FLAIR. PD and T1CE are excluded from the target inputs, making the external test modality matched to training. This experiment assesses persistence of the evaluation pattern under dataset/modality shift; it is not an external three-architecture ranking and is not a replacement for the primary five-modal MSSEG-1 comparison.

4.5. Evaluation, Uncertainty, and Reproducibility

Predictions are evaluated in the deterministic E1–E7 sequence in Section 3. Statistical procedures follow Section 3.9. We report voxel Dice/HD95/ASSD/PPV; lesion recall/precision/F1, FP/FN per scan, matched Dice/HD95 and volume error; exact topographic pattern and DIS-proxy agreement; and cluster type/Dice/HD95/volume error. Subject-mean and pooled-lesion quantities are labeled explicitly.

5. Results and Discussion

We first report uncertainty at subject level, then use pooled-lesion analyses to explain where the subject-level trade-offs originate. The central result is not a universal architecture ordering: it is that near-tied Dice does not imply equivalent lesion detection, false-positive burden, location preservation, or topology.

5.1. Near-Tied Dice, Separated Non-Overlap Endpoints

Table 2 reports subject means and 95% CIs; selected paired contrasts are detailed in Appendix Table 9. Dice intervals overlap strongly. In paired analysis, the SegResNet–nnU-Net Dice difference is 0.001 [–0.023, 0.031]. In contrast, nnU-Net improves lesion recall over SegResNet by 0.051 [0.016, 0.088], while SwinUNETR produces 10.67 more false-positive lesions per scan than nnU-Net [5.88, 16.38]. The exact-pattern difference between nnU-Net and SwinUNETR is 0.250 [0.000, 0.500]. Thus, the result is an endpoint-dependent preference with quantified trade-offs, not evidence that one model is universally superior.

The endpoint leaders consequently differ. SegResNet has the highest Dice point estimate, nnU-Net has the highest lesion recall point estimate and exact-pattern agreement, and SwinUNETR has the highest topography-aware recall (0.687 versus 0.647 for nnU-Net

Table 2: Subject-mean performance with 95% CIs on the 24-subject MSSEG-1 evaluation cohort. Dice, lesion, and FP intervals use subject bootstrap; binary topography intervals are exact Clopper–Pearson intervals.

Model	Voxel Dice	Lesion recall	Lesion precision	Lesion F1	FP/scan	Exact pattern	DIS-proxy agreement
nnU-Net	0.713 [0.648,0.772]	0.783 [0.724,0.842]	0.808 [0.743,0.866]	0.773 [0.733,0.811]	4.71 [3.75,5.75]	0.833 [0.626,0.953]	0.917 [0.730,0.990]
SegResNet	0.713 [0.669,0.759]	0.733 [0.653,0.810]	0.794 [0.724,0.857]	0.728 [0.679,0.776]	6.25 [4.38,8.33]	0.708 [0.489,0.874]	0.917 [0.730,0.990]
SwinUNETR	0.708 [0.656,0.756]	0.781 [0.712,0.844]	0.663 [0.574,0.747]	0.683 [0.624,0.746]	15.38 [10.25,21.25]	0.583 [0.366,0.779]	0.917 [0.730,0.990]

Table 3: Pooled lesion recall by size with subject-cluster bootstrap 95% CIs. GT denominators are shared across models.

Model	Small ($n = 194$)	Medium ($n = 708$)	Large ($n = 270$)
nnU-Net	0.175 [0.078,0.286]	0.758 [0.677,0.837]	0.963 [0.936,0.982]
SegResNet	0.180 [0.085,0.281]	0.692 [0.592,0.800]	0.944 [0.910,0.975]
SwinUNETR	0.304 [0.164,0.425]	0.749 [0.648,0.846]	0.970 [0.945,0.991]

and 0.629 for SegResNet), but at a clear precision/FP cost. A leave-one-subject-out jackknife preserves the full composite order in 19/24 iterations and changes it in 5/24. We therefore interpret the stable finding as absence of a Pareto-dominant model and endpoint dependence, not a fixed overall ranking.

5.2. Lesion Size Explains a Major Detection Mechanism

Pooled size-stratified recall (Table 3) is low for small lesions and near-saturated for large lesions. Because each voxel is 1 mm^3, these rows are physical-volume strata. Subject-cluster CIs are wide where lesion counts are concentrated in a few subjects, which appropriately limits comparative claims.

The full-cohort cross-model detection trace makes the mechanism concrete (Figure 2E). Of 1,172 GT lesions, 708 are detected by all three models, 238 are missed by all three, and 226 are model dependent (112 detected by exactly two and 114 by exactly one). Missed-by-all lesions are dominated by small/medium JC/cortical or IT components. In the pooled lesion-detection model, log lesion volume is the dominant predictor (odds ratio 3.86 per log-volume unit [3.45, 4.74]); increasing distance from ventricle and cortex is also associated with lower detection (OR 0.965 [0.947, 0.980] and 0.883 [0.828, 0.928] per mm). This identifies where evidence is lost rather than restating that Dice is incomplete.

5.3. Location-Specific Detection Cannot Be Inferred from Dice

The cohort contains 189 PV, 924 JC/cortical, and 59 IT lesions; IT is present in 12/24 subjects. Table 4 therefore presents a cohort/model finding, not a universal anatomical hierarchy. IT recall is lower than PV recall for every model. SwinUNETR attains the highest IT recall (0.458) but only 0.190 precision, illustrating why sensitivity and false-positive burden must be read together. Appendix Table 8 shows that volume-weighted PV capture exceeds 0.99 for all models, whereas IT capture is 0.719–0.789. These results are compatible with both anatomical difficulty and class imbalance; we do not claim the hierarchy would persist under IT-enriched training.

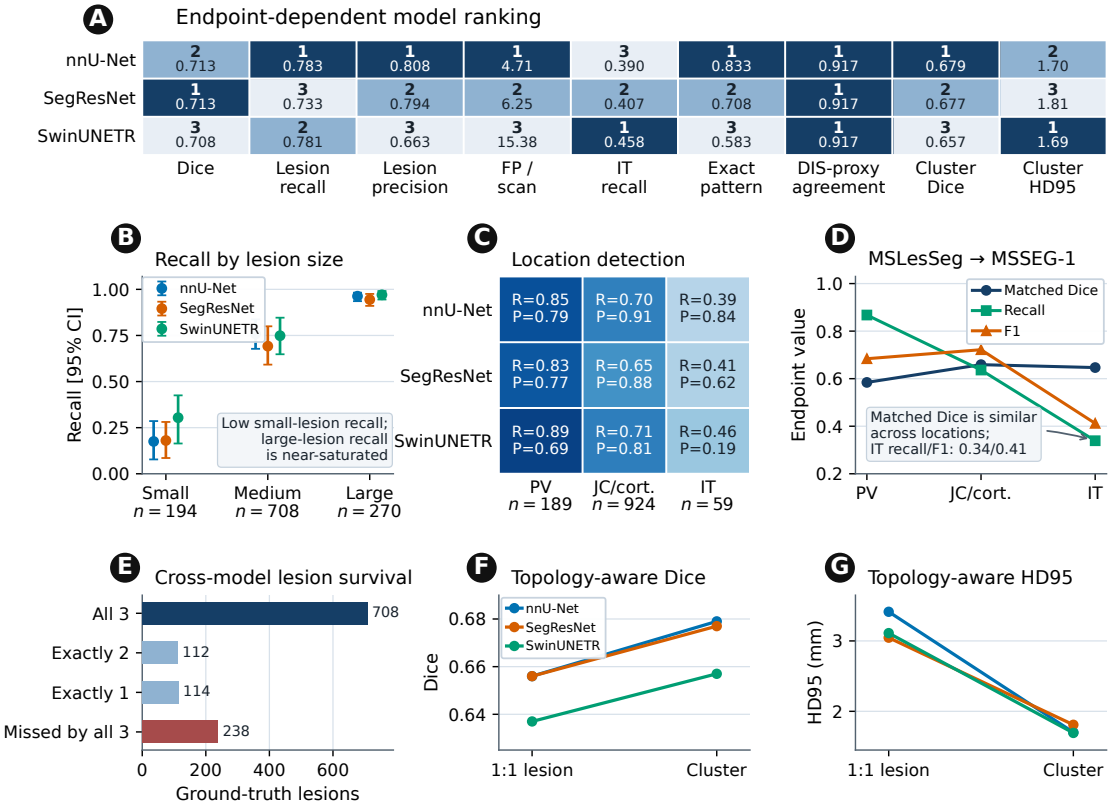

Figure 2: Compact summary of the clinically structured evaluation. (A) Endpoint-dependent ranking; each cell reports rank and value. (B) Size-stratified recall with subject-cluster bootstrap 95% CIs. (C) Location-stratified recall (R) and precision (P). (D) Modality-matched MSLesSeg-to-MSSEG-1 location endpoints. (E) Full-cohort lesion survival across models. (F–G) Recalibration from one-to-one lesion matching to topology-aware cluster scoring.

Anatomical-label sensitivity: FreeSurfer and SAMSEG assign the same primary class to 888/1,172 lesions (75.8%, Cohen's $\kappa = 0.542$); disagreements concentrate in small or multiregion boundary lesions. Crucially, subject-level DIS-proxy status agrees for all 24 subjects across anatomical engines. A 0.5–5% minimum-contact sweep changes location recall by at most 0.032 relative to the primary contact rule. These checks support robustness of the subject-level conclusion while making clear that lesion-level boundary labels are not perfectly interchangeable.

5.4. Subject Evidence, Component Filtering, and Topology

All models achieve DIS-proxy agreement 0.917, but exact location-pattern agreement ranges from 0.583 to 0.833 (Table 2). Each model produces two DIS-proxy false positives and no false negatives in this cohort. The proxy is intentionally narrow: it tests whether at least

Table 4: Pooled location-stratified lesion performance with subject-cluster bootstrap 95% CIs.

Model	Location (n)	Recall [95% CI]	Precision [95% CI]	F1
nnU-Net	PV (189)	0.852 [0.787,0.912]	0.790 [0.704,0.862]	0.820
	JC/cortical (924)	0.700 [0.631,0.776]	0.911 [0.861,0.946]	0.792
	IT (59)	0.390 [0.250,0.589]	0.839 [0.739,0.952]	0.532
SegResNet	PV (189)	0.831 [0.727,0.913]	0.768 [0.665,0.849]	0.798
	JC/cortical (924)	0.648 [0.559,0.746]	0.883 [0.820,0.933]	0.748
	IT (59)	0.407 [0.231,0.640]	0.619 [0.389,0.786]	0.491
SwinUNETR	PV (189)	0.894 [0.830,0.947]	0.693 [0.582,0.795]	0.781
	JC/cortical (924)	0.709 [0.619,0.788]	0.812 [0.740,0.867]	0.757
	IT (59)	0.458 [0.294,0.696]	0.190 [0.088,0.343]	0.269

Table 5: One-to-one matched versus topology-aware cluster geometry.

	Dice		HD95 (mm)	
Model	1:1	Cluster	1:1	Cluster
nnU-Net	0.656	0.679	3.414	1.702
SegResNet	0.656	0.677	3.046	1.809
SwinUNETR	0.637	0.657	3.111	1.693

two PV/JC/IT groups are represented and is not a clinical diagnosis or a replacement for longitudinal dissemination-in-time assessment.

Component filtering shows why reporting only the filtered Dice is unsafe. Removing components up to 10 mm^3 barely changes voxel Dice while jointly lowering lesion recall and FP burden; for nnU-Net, subject-mean recall decreases from 0.78 to 0.66 while FP/scan decreases from 4.71 to 2.62. Filtering is therefore a deployment trade-off, not a neutral cleanup step.

Topology-aware clusters consistently recalibrate geometric error (Figure 2F–G; Table 5). For nnU-Net, cluster Dice increases from 0.656 to 0.679 and HD95 decreases from 3.414 to 1.702; analogous improvements occur for SegResNet and SwinUNETR. The paired nnU-Net–SwinUNETR cluster-Dice difference is 0.024 [0.001, 0.048], whereas nnU-Net–SegResNet is 0.000 [–0.023, 0.021]. Cluster scoring does not excuse detection errors: GT-only and prediction-only clusters remain explicit. It separates complete misses and false positives from one-to-many, many-to-one, and many-to-many representation errors.

5.5. Independent MSLesSeg Validation

Table 6 reports experiments added after the initial submission. Same-domain MSLesSeg fold 0 uses 74 training and 19 validation scans. The external setting evaluates the MSLesSeg-trained three-modal model on three-modal MSSEG-1. Its lower Dice under dataset shift is not compared directly with the primary five-modal models; the experiment tests whether endpoint differences persist.

The external location analysis separates matched geometry from detection (Figure 2D). Matched-lesion Dice is similar across PV (0.584), JC (0.659), and IT (0.647). However, IT recall/F1 is 0.339/0.412 versus 0.868/0.684 for PV and 0.636/0.722 for JC. Once matched,

Table 6: Additional MS-domain validation. External CIs use subject bootstrap. Internal values are point estimates on the 19-scan fold-0 validation split.

Setting	Dice	Lesion recall	Small/large recall	DIS-proxy agreement
MSLesSeg to MSSEG-1	0.636 [0.547,0.710]	0.753 [0.678,0.823]	0.268/0.971	0.917
MSLesSeg fold 0	0.763	0.833	0.375/0.948	1.000

IT overlap is not markedly worse; the larger deficit is detection and false-positive control, which Dice cannot reveal.

5.6. Scope and Limitations

The 24-subject MSSEG-1 evaluation is a controlled paired benchmark, not a broad architecture leaderboard; single-model MSLesSeg testing cannot establish external rank stability. Location labels remain rule based and boundary sensitive, although subject-level DIS-proxy status and principal findings are stable in sensitivity checks. The proxy excludes dissemination in time, diagnosis, treatment response, and prospective impact. Rare E5 groups are descriptive; comparisons use full-cohort, subject-cluster, and independent MS analyses.

6. Conclusion

Beyond-Dice tests whether MS segmentation preserves clinical evidence. On 24 MSSEG-1 subjects, Dice-tied architectures differ in lesion recall, false-positive burden, topographic pattern, and topology-aware geometry. Of 1,172 lesions, 238 are missed by all models and volume strongly predicts detection; modality-matched MSLesSeg validation reproduces the separation between matched geometry and location-aware detection. PV/JC/IT labels and the DIS proxy are MS-specific, and this is not prospective validation. The broader principle is to match metrics to clinical evidence units, retain subjects as sampling units, report uncertainty and false-positive trade-offs, and audit rare presentations. Thus, "best model" is endpoint specific rather than a single-score label.

Data availability: MSSEG-1 and MSLesSeg were accessed under their respective data-use procedures.

Acknowledgments

This research was partially funded by the National Multiple Sclerosis Society (NMSS) under the "LAMINATE: Longitudinal AI-based MS Lesion and Atrophy Segmentation Tool" project.

References

S. Aslani et al. Multi-branch convolutional neural network for multiple sclerosis lesion segmentation. *NeuroImage*, 196:1–15, 2019. doi: 10.1016/j.neuroimage.2019.03.068.

Liviu Badea et al. Toward generalizable multiple sclerosis lesion segmentation models. *IEEE Access*, 13:97859–97869, 2025. doi: 10.1109/access.2025.3576145.

T. Brosch et al. Deep 3d convolutional encoder networks with shortcuts for multiscale feature integration applied to multiple sclerosis lesion segmentation. *IEEE Transactions on Medical Imaging*, 35(5):1229–1239, 2016. doi: 10.1109/tmi.2016.2528821.

A. Carass et al. Longitudinal multiple sclerosis lesion segmentation: Resource and challenge. *NeuroImage*, 148:77–102, 2017. doi: 10.1016/j.neuroimage.2016.12.064.

C. J. Clopper and Egon S. Pearson. The use of confidence or fiducial limits illustrated in the case of the binomial. *Biometrika*, 26(4):404–413, 1934. doi: 10.1093/biomet/26.4.404.

O. Commowick et al. Objective evaluation of multiple sclerosis lesion segmentation using a data management and processing infrastructure. *Scientific Reports*, 8(1):13650, 2018. doi: 10.1038/s41598-018-31911-7.

A. Danelakis et al. Survey of automated multiple sclerosis lesion segmentation techniques on magnetic resonance imaging. *Computerized Medical Imaging and Graphics*, 70:83–100, 2018. doi: 10.1016/j.compmedimag.2018.10.002.

Bradley Efron and Robert J. Tibshirani. *An Introduction to the Bootstrap*. Chapman and Hall, New York, 1993. doi: 10.1007/978-1-4899-4541-9.

Christine Egger et al. Mri flair lesion segmentation in multiple sclerosis: Does automated segmentation hold up with manual annotation? *NeuroImage: Clinical*, 13:264–270, 2017. doi: 10.1016/j.nicl.2016.11.020.

M. J. Fartaria et al. Partial volume-aware assessment of multiple sclerosis lesions. *NeuroImage: Clinical*, 18:245–253, 2018. doi: 10.1016/j.nicl.2018.01.011.

Bruce Fischl. Freesurfer. *NeuroImage*, 62(2):774–781, 2012. doi: 10.1016/j.neuroimage.2012.01.021.

D. García-Lorenzo et al. Review of automatic segmentation methods of multiple sclerosis white matter lesions on conventional magnetic resonance imaging. *Medical Image Analysis*, 17(1):1–18, 2013. doi: 10.1016/j.media.2012.09.004.

Nazem Ghasemi, Shahnaz Razavi, and Elham Nikzad. Multiple sclerosis: pathogenesis, symptoms, diagnoses and cell-based therapy. *Cell Journal*, 19(1):1–10, 2017. doi: 10.22074/cellj.2016.4867.

Francesco Guarnera et al. Mslesseg: baseline and benchmarking of a new multiple sclerosis lesion segmentation dataset. *Scientific Data*, 12(1):920, 2025. doi: 10.1038/s41597-025-05250-y.

Maha Haki et al. Review of multiple sclerosis: Epidemiology, etiology, pathophysiology, and treatment. *Medicine*, 103(8):e37297, 2024. doi: 10.1097/md.0000000000037297.

Ali Hatamizadeh et al. Swin unetr: Swin transformers for semantic segmentation of brain tumors in mri images, 2022.

A. Hindsholm et al. Scanner agnostic large-scale evaluation of ms lesion delineation tool for clinical mri. *Frontiers in Neuroscience*, 17:1177540, 2023. doi: 10.3389/fnins.2023.1177540.

Sebastian Hitziger et al. Triplanar u-net with lesion-wise voting for the segmentation of new lesions on longitudinal mri studies. *Frontiers in Neuroscience*, 16:964250, 2022. doi: 10.3389/fnins.2022.964250.

Fabian Isensee et al. nnu-net: Self-adapting framework for u-net-based medical image segmentation, 2018.

Saurabh Jain et al. Two time point ms lesion segmentation in brain mri: An expectation-maximization framework. *Frontiers in Neuroscience*, 10:576, 2016. doi: 10.3389/fnins.2016.00576.

B. Johnston et al. Segmentation of multiple sclerosis lesions in intensity corrected multispectral mri. *IEEE Transactions on Medical Imaging*, 15(2):154–169, 1996. doi: 10.1109/42.491417.

Amrita Kaur et al. State-of-the-art segmentation techniques and future directions for multiple sclerosis brain lesions. *Archives of Computational Methods in Engineering*, 28(3): 951–977, 2020. doi: 10.1007/s11831-020-09403-7.

Harold W. Kuhn. The hungarian method for the assignment problem. *Naval Research Logistics Quarterly*, 2(1–2):83–97, 1955. doi: 10.1002/nav.3800020109.

X. Lladó et al. Segmentation of multiple sclerosis lesions in brain mri: A review of automated approaches. *Information Sciences*, 186(1):164–185, 2012. doi: 10.1016/j.ins.2011.10.011.

Yang Ma et al. Multiple sclerosis lesion analysis in brain magnetic resonance images: Techniques and clinical applications. *IEEE Journal of Biomedical and Health Informatics*, 26 (6):2680–2692, 2022. doi: 10.1109/jbhi.2022.3151741.

Lena Maier-Hein, Annika Reinke, Patrick Godau, et al. Metrics reloaded: Recommendations for image analysis validation. *Nature Methods*, 21(2):195–212, 2024. doi: 10.1038/s41592-023-02151-z.

Richard McKinley et al. Simultaneous lesion and neuroanatomy segmentation in multiple sclerosis using deep neural networks, 2019.

Richard McKinley et al. Simultaneous lesion and brain segmentation in multiple sclerosis using deep neural networks. *Scientific Reports*, 11(1):1087, 2021. doi: 10.1038/s41598-020-79925-4.

Andriy Myronenko. 3d mri brain tumor segmentation using autoencoder regularization, 2018.

Andria Nicolaou et al. A systematic review of quantitative mri brain analysis studies in multiple sclerosis disease. *IEEE Access*, 12:162359–162391, 2024. doi: 10.1109/access.2024.3489798.

Oula Puonti, Juan Eugenio Iglesias, and Koen Van Leemput. Fast and sequence-adaptive whole-brain segmentation using parametric bayesian modeling. *NeuroImage*, 143:235–249, 2016. doi: 10.1016/j.neuroimage.2016.09.011.

Alessia Rondinella et al. Icpr 2024 competition on multiple sclerosis lesion segmentation - methods and results. In *Pattern Recognition. Competitions*, pages 1–16, 2024. doi: 10.1007/978-3-031-80139-6_1.

Alessandro Pasquale De Rosa et al. Consensus of algorithms for lesion segmentation in brain mri studies of multiple sclerosis. *Scientific Reports*, 14(1):21348, 2024. doi: 10.1038/s41598-024-72649-9.

M. Styner et al. 3d segmentation in the clinic: A grand challenge ii: Ms lesion segmentation. *The MIDAS Journal*, 2008. doi: 10.54294/lmkqvm.

S. Valverde et al. Improving automated multiple sclerosis lesion segmentation with a cascaded 3d convolutional neural network approach. *NeuroImage*, 155:159–168, 2017. doi: 10.1016/j.neuroimage.2017.04.034.

David R. van Nederpelt et al. Scanner-specific optimisation of automated lesion segmentation in ms. *NeuroImage: Clinical*, 44:103680, 2024. doi: 10.1016/j.nicl.2024.103680.

Liqun Wang et al. Survey of the distribution of lesion size in multiple sclerosis: implication for the measurement of total lesion load. *Journal of Neurology, Neurosurgery & Psychiatry*, 63(4):452–455, 1997. doi: 10.1136/jnnp.63.4.452.

Simon K. Warfield, Kelly H. Zou, and William M. Wells. Simultaneous truth and performance level estimation (STAPLE): An algorithm for the validation of image segmentation. *IEEE Transactions on Medical Imaging*, 23(7):903–921, 2004. doi: 10.1109/TMI.2004.828354.

Mike P Wattjes et al. MAGNIMS consensus guidelines on the use of MRI in multiple sclerosis—establishing disease prognosis and monitoring patients. *Nature Reviews Neurology*, 11(10):597–606, 2015. doi: 10.1038/nrneurol.2015.157.

Yicheng Wu et al. Coactseg: Learning from heterogeneous data for new multiple sclerosis lesion segmentation. In *Medical Image Computing and Computer Assisted Intervention – MICCAI 2023*, pages 3–13, 2023. doi: 10.1007/978-3-031-43993-3_1.

Chenyi Zeng et al. Review of deep learning approaches for the segmentation of multiple sclerosis lesions on brain mri. *Frontiers in Neuroinformatics*, 14:610967, 2020. doi: 10.3389/fninf.2020.610967.

Hang Zhang et al. All-net: Anatomical information lesion-wise loss function integrated into neural network for multiple sclerosis lesion segmentation. *NeuroImage: Clinical*, 32:102854, 2021. doi: 10.1016/j.nicl.2021.102854.

Appendix A. Reproducibility Details

A.1. Training and Inference Configuration

Table 7: Training and inference configuration. "Auto" denotes values generated by the nnU-Net v2 plan rather than manually selected values.

Model	Loss/optimizer	Learning rate/schedule	ROI; batch	Sampling	Regularization	Architecture	Inference
nnU-Net v2, 3D full resolution	Dice+CE; SGD, momentum 0.99	10^{-2}; polynomial default	Auto; auto	foreground oversampling 0.33	weight decay 3×10^{-5}	default trainer and generated plan	folds 0–4, native probability ensemble
SegResNet	DiceCE; AdamW	2×10^{-4}; cosine	128^3; 1	4 samples/case	decay 10^{-5}; dropout 0.2	32 filters; down [1,2,2,4], up [1,1,1]	softmax fold mean; 0.5 sliding overlap
SwinUNETR	DiceCE; AdamW	10^{-4}; cosine	96^3; 1	2 samples/case	decay 10^{-5}; checkpointing	feature 24; depths [2,2,2,2]; heads [3,6,12,24]	softmax fold mean; 0.5 sliding overlap

A.2. Statistical Estimands

Subject-level tables estimate the mean performance on a cohort of subjects. A bootstrap replicate samples 24 subject identifiers with replacement and averages their per-subject metrics. Paired model differences use the same sampled identifiers for both models. Lesion-level tables estimate pooled lesion recall or precision; bootstrap replicates sample subjects and include every lesion belonging to each sampled subject. This preserves within-subject clustering and allows a subject with many lesions to contribute its complete lesion set rather than treating lesions as independent patients. Binary subject proportions use exact Clopper–Pearson limits.

The reported random seed is 20260606 and the number of resamples is 5,000. The leave-one-subject-out audit repeats the composite endpoint ranking 24 times. It preserves the baseline order in 19 iterations and produces a different order in 5, spanning three unique orders. We therefore report endpoint dependence and no Pareto-dominant model as the robust claim.

A.3. Location QC and Sensitivity

For each GT and predicted component, the analysis QC row contains: subject and lesion identifier, volume in voxels and mm^3, primary and secondary location, distance to ventricle and cortex in mm, PV/JC/IT overlap voxel counts and fractions, binary contact indicators, a multiregion flag, and the anatomy engine. The main assignment is PV > IT > JC/cortical > other after 3 mm dilation.

The minimum-contact analysis replaces "overlap > 0" with overlap fraction thresholds 0.5%, 1%, 2%, and 5%, while retaining the priority order. Across these thresholds, the largest absolute change in a main location-recall endpoint is at most 0.032. The independent anatomy-engine analysis compares primary labels component by component: 888/1,172 agree (75.8%, $\kappa = 0.542$), and the derived subject-level DIS-proxy status is identical for all 24 subjects. These results support robustness of the main endpoint direction while documenting residual boundary ambiguity.

A.4. Rater Uncertainty

MSSEG-1 evaluation uses the supplied consensus reference. We therefore evaluate against one consensus mask rather than treating seven expert masks as independent observations. Where individual masks are available in future cohorts, the same endpoint pipeline can be repeated per rater or applied to a majority/STAPLE-style probabilistic consensus (Warfield et al., 2004); this is an extension, not an analysis claimed here.

Appendix B. Additional Result Tables

B.1. Location Volume Capture

Table 8: Pooled count recall versus volume-weighted recall by location.

Model	Location	Count recall	Volume-weighted recall
nnU-Net	PV	0.852	0.994
	JC/cortical	0.700	0.909
	IT	0.390	0.749
SegResNet	PV	0.831	0.991
	JC/cortical	0.648	0.883
	IT	0.407	0.719
SwinUNETR	PV	0.894	0.996
	JC/cortical	0.709	0.907
	IT	0.458	0.789

B.2. Paired Model Differences

Table 9: Selected paired subject-bootstrap differences. Positive values favor the first model for recall/pattern/Dice and indicate greater burden for FP/scan.

Comparison	Endpoint	Difference [95% CI]
SegResNet – nnU-Net	Voxel Dice	0.001 [–0.023,0.031]
nnU-Net – SegResNet	Lesion recall	0.051 [0.016,0.088]
nnU-Net – SwinUNETR	Lesion recall	0.003 [–0.027,0.031]
SwinUNETR – nnU-Net	FP/scan	10.67 [5.88,16.38]
nnU-Net – SwinUNETR	Exact pattern	0.250 [0.000,0.500]
nnU-Net – SwinUNETR	Cluster Dice	0.024 [0.001,0.048]
nnU-Net – SegResNet	Cluster Dice	0.000 [–0.023,0.021]

Appendix C. Clinical Interpretation Boundaries

The DIS proxy is

$$\text{DIS-proxy} = \mathbf{1}\{|\{\text{PV}, \text{JC/cortical}, \text{IT}\}| \geq 2\}\,. \tag{5}$$

It measures preservation of a spatial evidence pattern under the paper's lesion-labeling rules. It does not encode lesion morphology, symptomatic status, spinal-cord involvement,

dissemination in time, differential diagnosis, or treatment response. Accordingly, "DIS-proxy false positive" means that the predicted segmentation contains at least two relevant location groups when the reference does not; it is not itself a false clinical diagnosis.

The terms screening and monitoring are used only to illustrate endpoint priorities. A sensitivity-oriented review queue may tolerate more candidate lesions, whereas longitudinal quantification may prioritize precision and stable topology. Prospective studies are required to establish thresholds, workflow utility, reader interaction, and patient outcomes.

Appendix D. Qualitative Failure-Mode Analysis

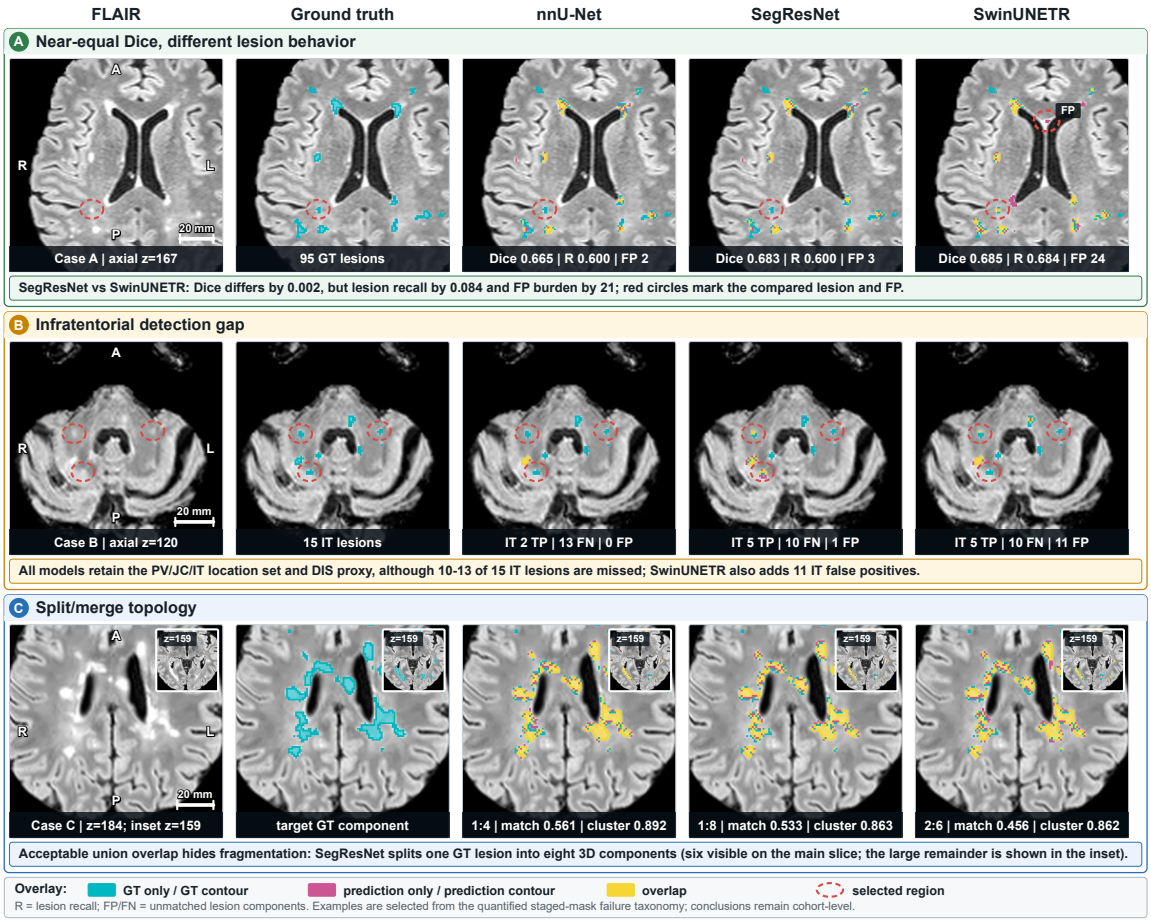

Figure 3: Qualitative failure modes. (A) Near-equal Dice masks differences in lesion recall and false-positive (FP) burden. (B) All models preserve the PV/JC/IT location set and DIS proxy despite missing 10–13/15 IT lesions. (C) Overlap scores obscure split/merge fragmentation; SegResNet splits one ground-truth lesion into eight components. Values are case- or cluster-level as labeled.