REFERENCE-DEPENDENT SCORES, MARGIN-STABLE DECISIONS: CALIBRATING MRI RECONSTRUCTION BENCHMARKS

Anonymous authors
Paper under double-blind review

ABSTRACT

Reference-based MRI benchmarks conflate reconstruction quality with the statistical compatibility between an input and its evaluation target. We introduce a calibration framework for this dependence in repeated-acquisition MRI. When a reference contains the same measured repetition used to form an accelerated input, realization-specific reconstruction error can correlate with reference noise and shift method scores even when repeat count and averaging are matched. We separate this method-level score distortion from the downstream decision instability it may induce. A margin-normalized certificate, ρ, makes the distinction explicit: the disjoint-reference winner is guaranteed to remain the winner when $\rho < 1$, whereas near-tied candidates are vulnerable to much smaller differential score shifts. On M4Raw 0.3T data, an all-slice T1/T2 study over 540 subject–contrast–slice units yields positive independent-validator loss for shared-reference selection (0.00916 NRMSE on V_1 and 0.01389 on V_2), and the direction transfers to 270 FLAIR units. Adding the official PI-VarNet nearly removes selection harm because the learned model creates a large candidate margin, even though its shared–disjoint score shift is the largest in the pool. A second competitive learned candidate restores 70/540 per-unit selector changes while independent-validator loss remains small. Within the classical pool, a runner-up margin approximation explains 205 of 220 observed selector changes with no false positives. These results motivate a practical reporting protocol for repeated-acquisition benchmarks: pair reference provenance and score sensitivity with candidate-margin-based decision stability.

1 INTRODUCTION

MRI reconstruction benchmarks usually treat the reference image as a fixed endpoint: reconstruct an accelerated input, compare it with a target, and use the resulting NRMSE, PSNR, or SSIM to rank methods. In repeated or retrospectively undersampled acquisitions, however, the target can share measured data with the reconstruction input. A method that preserves realization-specific structure from the input may then match an overlapping reference unusually well, even if that advantage does not transfer to a fresh acquisition of the same anatomy.

This issue is distinct from reference noise. Two references can contain the same number of repetitions, use the same averaging rule, and have similar nominal SNR, yet differ in whether one of those repetitions is also the reconstruction input. For a reconstruction $\hat{x}_m = x + e_m$ and a measured reference $R = x + n_R$, squared error contains the cross term $-2\mathbb{E}\langle e_m, n_R\rangle$. Reusing the input acquisition can make this term method-dependent. An independent reference removes that direct overlap but remains noisy; it is not assumed to be ground truth.

We study this distinction on M4Raw (Lyu et al., 2023), a public 0.3T dataset with raw multi-coil k-space and repeated T1, T2, and FLAIR acquisitions. Matched references let us change overlap while holding repeat count fixed, and later repetitions provide validators that are disjoint from both selector references. The expanded study covers all valid slices in a participant-disjoint 15-subject cohort (540 T1/T2 units), a third contrast (270 FLAIR units), classical reconstruction pools of different strength, the official M4Raw PI-VarNet, and a separate lightweight learned reconstructor selected using development data only.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

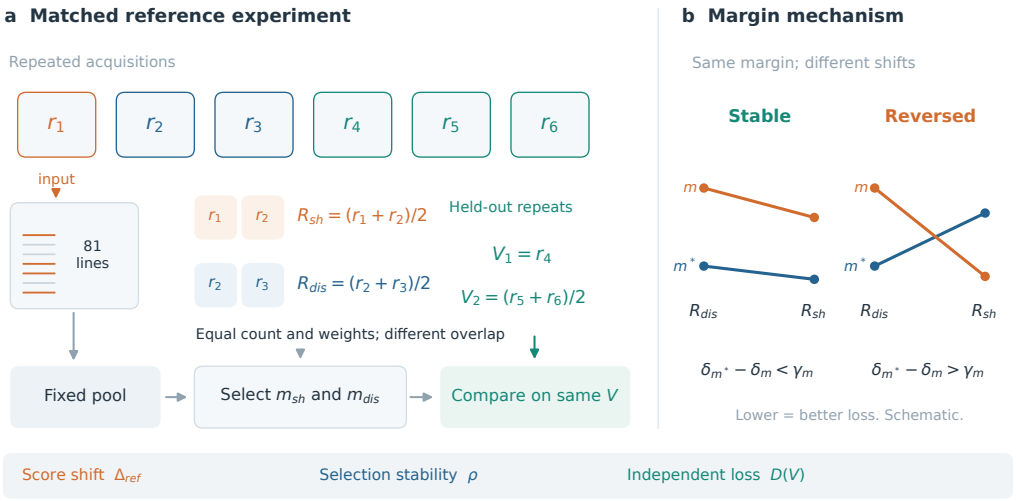

Figure 1: Matched reference experiment and margin mechanism. (a) Reconstruct the same masked input from r_1 once, then select from a fixed pool using equally averaged shared and disjoint selector references; assess both choices on the same independent validators. (b) Schematic score paths show that absolute score shifts can preserve or reverse the winner depending on the differential shift relative to the candidate margin. The right panel is illustrative, not measured data.

The enlarged experiment changes the interpretation in a useful way. In classical pools, overlap changes method selection and the shared-reference choice performs worse on independent measurements. Adding PI-VarNet alone nearly removes selection changes because it dominates the candidate pool. Adding an independent competitive learned reconstruction then restores 70/540 per-unit selection changes. The independent-validator penalty remains small, however, showing that candidate competition determines whether a score shift changes a decision while the selected methods' actual loss gap determines its downstream cost. We distinguish all three quantities rather than calling any one of them *bias*.

Our contributions are:

- We formalize **reference overlap** as an evaluation variable separate from reference SNR and distinguish method-level score distortion from downstream selection harm.
- We establish the decision effect at full-volume scale on T1/T2 (540 units) and on a third contrast, FLAIR (270 units), using subject-cluster confidence intervals and independent measured validators.
- We vary the candidate pool with one and two competitive learned reconstructors. Single-model dominance buffers selection changes, while restoring learned-method competition brings them back without producing a comparable independent-validator penalty.
- We connect these observations to benchmark design: whether overlap matters operationally depends on both reference construction and the selection margin among candidate methods.

2 Related Work

MRI reconstruction and reference-based benchmarks. Parallel imaging and compressed sensing established the modern inverse-problem view of accelerated MRI (Pruessmann et al., 1999; Griswold et al., 2002; Lustig et al., 2007; Uecker et al., 2014). Learned reconstruction subsequently combined data consistency with trainable priors, including deep cascades (Schlemper et al., 2018), cross-domain networks (Eo et al., 2018), variational networks (Hammernik et al., 2018), MoDL (Aggarwal et al., 2019), and end-to-end variational networks (Sriram et al., 2020). Large benchmarks such as fastMRI (Zbontar et al., 2018; Knoll et al., 2020) made reference-based metrics such as NRMSE and SSIM standard for model comparison. These metrics are useful, but their interpretation depends on what

the reference represents. Separate work on reconstruction instability and hidden preprocessing has likewise shown that seemingly minor evaluation or data-pipeline choices can produce misleading conclusions (Antun et al., 2020; Shimron et al., 2022).

Noisy targets and self-supervision. Noise2Noise (Lehtinen et al., 2018) and Noise2Self (Batson & Royer, 2019) show that independent noisy observations can support learning without clean targets under suitable assumptions. MRI-specific self-supervision similarly avoids fully sampled targets by splitting acquired measurements (Yaman et al., 2020) or leveraging repeated acquisitions. Rep2Rep (Janjusevic et al., 2026), for example, explicitly uses distinct low-field repetitions for noise-adaptive MRI denoising. These methods motivate the importance of independence, but they address training rather than ranking bias caused by an input-shared evaluation reference.

Reference noise and image-quality estimation. SURE and its Monte Carlo variants estimate risk without access to a clean target (Stein, 1981; Ramani et al., 2008); correlated-noise extensions have also been studied in learned denoising (Zhussip et al., 2019). Most directly related, Wang et al. (2024) demonstrated the "hidden noise" problem in MRI reconstruction: a conventional noisy reference can prefer different reconstructions from a higher-quality multi-average reference, and noise-aware metrics can mitigate the discrepancy. In their experimental design the lower-quality reference is also the acquisition from which the input is retrospectively undersampled, while the higher-quality reference uses more averages and additional denoising. We complement that result by holding the number and weight of repetitions fixed, changing only whether a selector reference contains the input acquisition, and then measuring the consequence of the resulting decision on separate validator repetitions.

3 Reference Dependence in Reconstruction Evaluation

3.1 A simple decomposition

Let the latent image be x, the reconstruction error of method m be e_m, and a measured reference be $R = x + n_R$. For squared error,

$$\mathbb{E}\|\hat{x}_m - R\|_2^2 = \mathbb{E}\|e_m\|_2^2 + \mathbb{E}\|n_R\|_2^2 - 2\,\mathbb{E}\langle e_m, n_R\rangle. \tag{1}$$

If R is independent of the input that generated e_m, the cross term is often small under standard zero-mean assumptions. If the reference reuses the same measured realization, however, input-correlated reconstruction residuals and n_R can share components. Two references with identical repeat count and nominal SNR can therefore induce different rankings through the covariance term in Eq. 1. Our experiments do not attempt to estimate this covariance directly. Instead, they intervene on overlap while keeping repeat count fixed, then test the selected method on independent measurements.

3.2 Matched references and selection

For an input reconstructed from repetition 1, define

$$R_{\mathrm{sh}} = \tfrac{1}{2}(r_1 + r_2), \qquad R_{\mathrm{dis}} = \tfrac{1}{2}(r_2 + r_3), \tag{2}$$

where r_j is the root-sum-of-squares magnitude image from a real measured repetition. Both references use two repetitions with weight $1/2$; only R_{sh} contains the input acquisition. For a method pool $\mathcal{M}$,

$$\hat{m}_{\mathrm{sh}} = \arg\min_{m\in\mathcal{M}} \mathrm{NRMSE}(\hat{x}_m, R_{\mathrm{sh}}), \quad \hat{m}_{\mathrm{dis}} = \arg\min_{m\in\mathcal{M}} \mathrm{NRMSE}(\hat{x}_m, R_{\mathrm{dis}}). \tag{3}$$

The primary decision-level estimand on an independent validator V is

$$D_{\mathrm{sel}}(V) = \mathbb{E}_u\big[\mathrm{NRMSE}(\hat{x}_{\hat{m}_{\mathrm{sh}}}, V) - \mathrm{NRMSE}(\hat{x}_{\hat{m}_{\mathrm{dis}}}, V)\big], \tag{4}$$

where u indexes balanced subject/contrast/slice units. Positive values mean selection using the shared reference performs worse on unseen measurements.

4 EXPERIMENTAL SETUP

4.1 DATASET AND REFERENCE CONSTRUCTION

We use M4Raw Brain V1.6 (Lyu et al., 2023), acquired at 0.3T with a four-channel head coil. T1 and T2 provide six repetitions per test participant; FLAIR provides four. The accelerated input is repetition r_1 under the frozen R=3 support used by the prior study (81 acquired phase-encoding lines for T1/T2). For T1/T2, the matched selector references are

$$R_{\mathrm{sh}} = \tfrac{1}{2}(r_1 + r_2), \qquad R_{\mathrm{dis}} = \tfrac{1}{2}(r_2 + r_3),$$

with independent validators $V_1 = r_4$ and $V_2 = (r_5 + r_6)/2$. For FLAIR, the same two-reference construction is used with the single available independent validator $V_1 = r_4$. NRMSE is primary; PSNR and SSIM are secondary agreement measures.

4.2 RECONSTRUCTION POOLS

The classical three-method pool contains zero-filled reconstruction (ZF), regularized SENSE, and a SENSE reconstruction followed by TV denoising. A five-method pool adds two pre-specified TV regularization settings, with the fixed configuration chosen before the confirmation cohort is opened. We call the three-classical pool C3 and the extended five-classical pool C5. Our learned candidates are the official M4Raw PI-VarNet checkpoint and M4c2, a separate 0.48M-parameter ZF-skip CNN with shape-guarded training and development-only checkpoint selection. Both consume the frozen 81-line input. We compare C3, C5, C3+VarNet, C3+M4c2 and C3+both; these are candidate-pool changes, not a new reconstruction-architecture claim. Preprocessing, train/test separation, and M4c2 capability tests appear in the supplement.

4.3 EVALUATION UNITS AND STATISTICS

The primary full-volume T1/T2 analysis uses the 15-subject confirmation cohort, two contrasts, and all 18 valid slices, for 540 units. FLAIR contributes 15 subjects $\times$ 18 slices = 270 units. Selection loss is

$$D(V) = L(\hat{x}_{m_{\mathrm{sh}}}, V) - L(\hat{x}_{m_{\mathrm{dis}}}, V),$$

where m_{sh} and m_{dis} are the methods selected by the shared and disjoint references. Positive D means shared-reference selection is worse on independent measurements. Confidence intervals use 20,000 bootstrap resamples clustered by subject. We also report the per-method reference shift

$$\Delta_{\mathrm{ref}}(m) = L(\hat{x}_m, R_{\mathrm{sh}}) - L(\hat{x}_m, R_{\mathrm{dis}}),$$

which measures score distortion even when both references select the same method.

5 RESULTS

5.1 FULL-VOLUME T1/T2 REPLICATION

The slice-restricted effect survives when every valid slice is included. In the three-method pool, the shared and disjoint references select different methods on 220/540 units (40.7%). Selection with the shared reference increases independent-validator NRMSE by **0.009159** (95% CI [0.00702, 0.01137]) on V_1 and **0.013887** [0.01066, 0.01730] on V_2. The effect is larger for T1 than T2 but positive for both contrasts (Table 1). The earlier two-slice cohorts therefore did not create the result; their effect sizes fall in the same range as the full-volume estimate.

5.2 THE EFFECT EXTENDS TO A THIRD CONTRAST

FLAIR provides four repetitions, allowing the same matched shared/disjoint selector construction with one independent validator. Across 270 units, the three-method pool has 37 selection changes and $D(V_1) = \mathbf{0.004236}$ with 95% CI [0.00192, 0.00699]. A five-method pool remains positive at 0.001892 [0.00056, 0.00355]. The FLAIR magnitude is smaller than T1/T2, but the direction is unchanged. This provides multi-contrast support without treating a single noisy validator as latent ground truth.

Table 1: Selection loss under matched reference construction. Positive values mean that the method selected by the input-shared reference performs worse on an independent validator. FLAIR has only one independent validator.

Setting	Pool	$D(V_1)$	$D(V_2)$
T1, all slices	3 classical	0.01329	0.02002
T2, all slices	3 classical	0.00502	0.00775
T1+T2 pooled	3 classical	**0.00916**	**0.01389**
T1+T2 pooled	5 classical	0.00726	0.01086
FLAIR	3 classical	**0.00424**	–
T1+T2 pooled	C3 + VarNet	0.00009	0.00013
T1+T2 pooled	C3 + M4c2	0.00000	0.00000
T1+T2 pooled	C3 + both	**0.00047**	**0.00062**

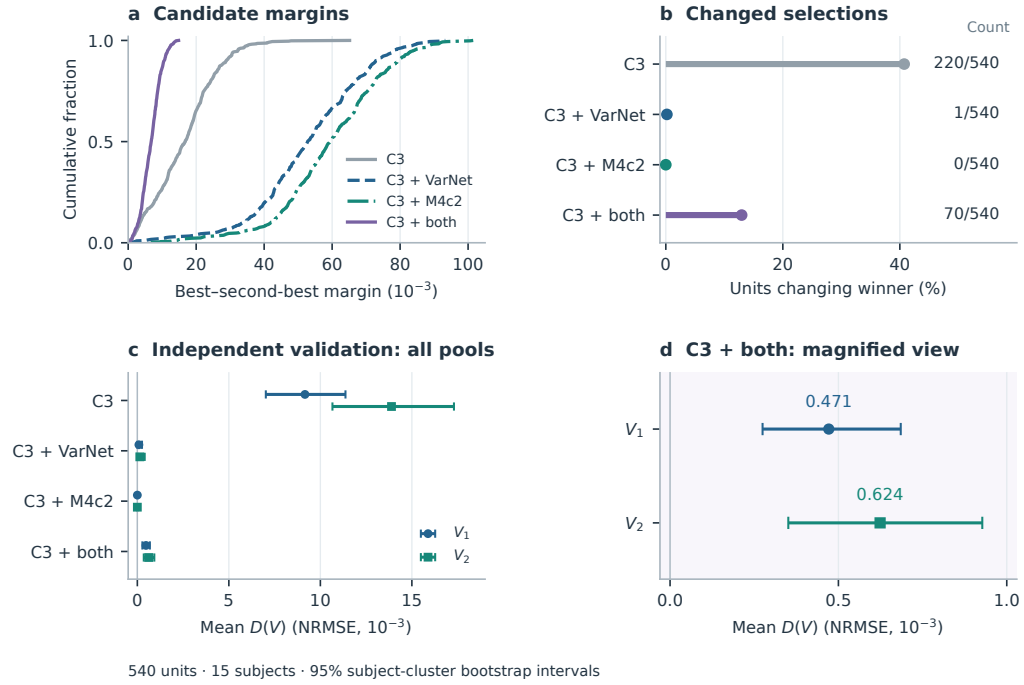

Figure 2: Pool composition modulates score-to-decision propagation across the same 540 T1/T2 units. (a) Disjoint-reference best–second margins. (b) Per-unit shared/disjoint winner changes. (c) Mean independent-validator loss $D(V)$ with subject-cluster 95% confidence intervals; note that C3 and learned pools have markedly different scales. (d) Magnified learned-pair result. C3+M4c2 is a single-dominant-model control; C3+both restores competition. Pool changes affect margins and differential score shifts jointly and do not isolate a causal margin-only intervention.

5.3 Candidate-pool competition separates score distortion from decision harm

In C3, selector references disagree on 220/540 units. With VarNet added, the selections disagree on only 1/540 units: its disjoint-reference median top-two margin is 0.05280, compared with 0.01679 for C3. C3+M4c2 alone is also stable (0/540 changes; median margin 0.05959). Both learned models are strong enough to dominate the corresponding four-method pools, but neither is independent of reference provenance at the score level.

Adding *both* learned candidates shrinks the median disjoint best–second margin to 0.00678 and restores **70/540 (13.0%)** changed per-unit selections, 66 on T1 and four on T2 (Fig. 2). For all 70 changing units, the shared selector chooses VarNet whereas the disjoint selector chooses M4c2. M4c2 remains the aggregate best fixed method under either reference; the result is *per-unit selection*

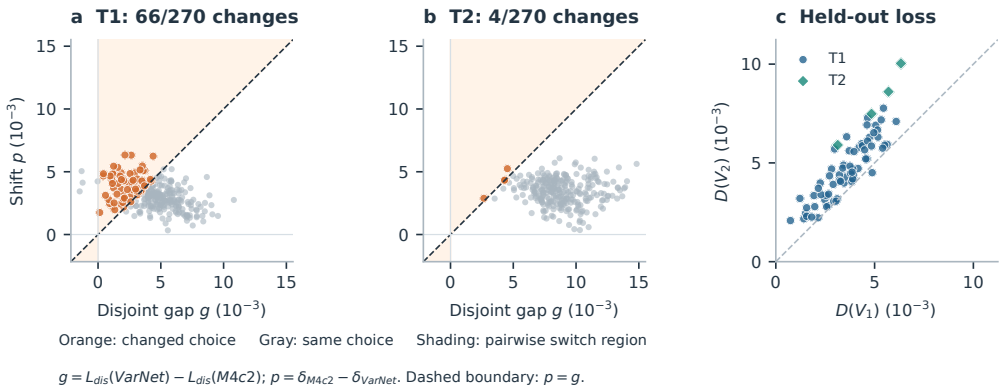

Figure 3: The two-learned-candidate comparison. (a–b) Disjoint signed score gap between VarNet and M4c2 versus their differential reference shift for T1 and T2; the schematic boundary partitions pairwise stable and switched orderings. Orange denotes changed *full-pool* selections. (c) Loss on independent validators among the 70 changed units only, shown separately from the whole-cohort losses in Fig. 2. This is a post-hoc explanation, not a prospective classifier of future scans.

instability, not a reversal of the aggregate leaderboard. The independent-validator penalty for C3+both is positive but small: $D(V_1) = 0.000471$ (95% CI [0.000275, 0.000685]) and $D(V_2) = 0.000624$ [0.000351, 0.000927]. The conditional loss among switching units is larger and must not be substituted for these whole-cohort means.

The per-method shared–disjoint score shift is nonzero for both learned models; Fig. S3 in the supplement includes their subject-cluster summaries. Hence a single dominant learned method does not remove reference dependence; a second competitive method reveals its decision-level effect again. Conversely, the fact that the renewed decision changes yield a small held-out penalty shows why *flip frequency* and *independent-validator loss* must be reported separately. Pool expansions jointly alter both candidate margins and method-specific reference shifts, so these comparisons test pool dependence without identifying margin as the sole causal mediator.

5.4 Acquisition localization supports the overlap interpretation

A four-corner analysis separates the 81 k-space lines used by the accelerated input from the other 114 acquired lines in the same repetition. Breaking overlap in the 81 input lines changes the TV–ZF gap by -0.021323 NRMSE (95% CI $[-0.022691, -0.019863]$), whereas the corresponding effect of the other 114 lines is only $+0.000797$ [0.000308, 0.001269]. Supplementary Fig. S4 summarizes the intervention. Additional rotation, swap, metric-agreement, and implementation checks are reported in the supplement rather than used as headline evidence.

6 Reference Construction as an Evaluation Design Choice

The experiments suggest a two-level view of reference dependence. At the *method level*, overlap changes the numerical score assigned to a fixed reconstruction. At the *decision level*, that score change matters only when it is large enough relative to the margin between competing methods to alter the winner. The single- and two-learned-candidate pools provide contrasting examples: score dependence persists in both, while decision changes track the competition between VarNet and M4c2.

This distinction explains why reference overlap can be scientifically important without producing a universal penalty. When candidates are close, realization-specific correlation can move their ordering; when one method dominates, the same bias may leave the decision unchanged. Evaluation sensitivity therefore depends jointly on reference construction, reconstruction error structure, and the competitive margin of the candidate pool.

Table 2: When reference overlap is most likely to affect a benchmark decision.

Factor	Lower sensitivity	Higher sensitivity
Measurement overlap	Disjoint acquisition data	Reference contains input measurements
Residual correlation	Methods suppress realization-specific structure similarly	Methods preserve shared structure to different degrees
Selection margin	One candidate clearly dominates	Several candidates are near-tied

For repeated-acquisition studies, a practical protocol follows: report which repetitions contribute to the reconstruction input and reference; if method or hyperparameter selection uses an overlapping reference, repeat the selection with a matched disjoint reference when possible; and evaluate both choices on measurements used by neither selector. When repeats are unavailable, the same principle motivates measurement splitting or noise-aware estimators that reduce direct reuse of the reconstruction input in model selection.

6.1 Selection stability depends on score margin

The learned-model result can be expressed with a simple stability condition. Let s_m^{dis} be the disjoint-reference score of method m and let $\delta_m = s_m^{\mathrm{sh}} - s_m^{\mathrm{dis}}$ denote the overlap-induced score shift. Suppose method $m^\star$ is best under the disjoint reference with margin

$$\gamma = \min_{m \neq m^\star} \left(s_m^{\mathrm{dis}} - s_{m^\star}^{\mathrm{dis}} \right) > 0.$$

The winner is unchanged under the shared reference whenever

$$\delta_{m^\star} - \delta_m < \gamma \qquad \text{for all } m \neq m^\star.$$

For a unique disjoint winner define $\gamma_m = s_m^{\mathrm{dis}} - s_{m^\star}^{\mathrm{dis}} > 0$ and the post-hoc normalized ratio

$$\rho = \max_{m \neq m^\star} \frac{\delta_{m^\star} - \delta_m}{\gamma_m}.$$

If $\rho < 1$ the disjoint winner remains strictly optimal; $\rho = 1$ permits a tie and $\rho > 1$ means some competitor overtakes it. The test uses both references and is a retrospective certificate, not a predictor available from a single reference. Thus a large absolute score distortion need not cause selection harm if the best method has a sufficiently large margin. Conversely, even modest differential distortion can change the winner in a near-tied pool. This criterion is algebraic; it describes both the buffered single-learned pools and the renewed instability with two learned candidates: score shifts remain visible even with a dominant learned candidate, whereas restoring a close competitor makes winner changes reappear.

For the 540 C3 units, restricting the test to the disjoint runner-up m_2 gives $\delta_{m^\star} - \delta_{m_2} > \gamma_{m_2}$. It flags 205 of the 220 observed winner changes and none of the 320 stable units; the full ρ criterion captures all 220 changes. This runner-up approximation uses the same observed score shifts as ρ.

This distinction is important for interpreting benchmark sensitivity. A benchmark can be biased at the score level without changing a headline leaderboard when one method is far ahead. The same benchmark may become decision-sensitive when tuning closely related checkpoints, regularization values, or architectures whose performance differs by less than the overlap-induced score shift. Hyperparameter selection can therefore be more vulnerable than coarse model-family comparison even when both use the same metric and reference.

Cohort consistency. The 10-subject exploratory and participant-disjoint 15-subject confirmation cohorts show effects consistent with the 540-unit all-slice extension; exact intervals, secondary metrics and the cohort comparison table are reported in the supplement.

6.2 What the learned extension changes scientifically

The learned-candidate experiment rules out an interpretation in which only classical reconstructions inherit the overlap effect. VarNet exhibits a large Δ_{ref}, yet it simultaneously makes downstream

Table 3: Common MRI reference constructions emphasize different evaluation questions.

Reference construction	What the score emphasizes	Main caveat
Same acquisition	Fidelity to the measured realization used for input formation	Direct input–reference dependence
Input + independent repeat(s)	Lower-variance target with partial realization sharing	Dependence decreases but does not vanish
Independent repeat(s)	Transfer to another measurement of the same anatomy	Inter-repeat noise, motion, physiology
Noise-aware risk estimate	Expected error under an explicit measurement/noise model	Validity depends on model assumptions

selection almost insensitive when it dominates the pool; a second competitive learned method restores decision sensitivity. This separates two questions that are often conflated: "Does the reference favor a method differently?" and "Does that favoritism change the selected method?" The first remains strongly positive for VarNet; the second is nearly null in the enlarged pool.

This distinction also changes how a negative result should be read on future datasets. If a strong learned model is selected under every reasonable reference construction, then a near-zero selection penalty does not demonstrate reference independence. One must inspect per-method score shifts or use a candidate set with enough competitive headroom to test whether rankings can move. Conversely, if both score shifts and selection changes are negligible, that would identify an operating regime in which overlap is practically unimportant.

6.3 Reference construction changes the estimand

The overlap question is not merely a recommendation to use "better" targets. Different constructions correspond to different estimands. A same-acquisition target asks how closely the reconstruction matches a realization from which its input samples were drawn. A disjoint repeat asks how well the reconstruction predicts another measurement of the same anatomy. A many-repeat average reduces variance but can still retain partial input dependence if the input repetition is included. Noise-aware risk estimators replace a measured target with a model-based quantity and introduce their own assumptions. These choices can all be useful, but they should not be treated as equivalent instances of "ground truth."

This perspective is particularly relevant to retrospective undersampling, where the unmasked acquisition is routinely treated as the reference for an input obtained by masking that same file. The construction is convenient and often scientifically appropriate, but its score should be interpreted as same-realization fidelity. If the goal is to choose a reconstruction that transfers to a fresh acquisition, an independent validator measures a different quantity. Our experiments quantify the gap between these two uses without claiming that one reference is universally superior.

7 Implications for Benchmark Design

The results suggest that reference construction should be treated as a benchmark design variable rather than as a preprocessing detail. At minimum, an MRI reconstruction benchmark should report whether the reference shares raw measurements with the evaluated input, how many repetitions contribute to the reference, and whether model or hyperparameter selection is performed against the same construction used for final reporting. These choices determine the statistical relation between reconstruction error and reference noise even when image dimensions, repeat count, and nominal SNR are matched.

The learned-pool results also motivate reporting both *method-level* and *decision-level* sensitivity when a benchmark is used for model selection. Per-method score shifts reveal whether a reference favors reconstructions differently, whereas selection loss additionally depends on the margins among candidate methods. VarNet alone nearly eliminates selection disagreement while retaining a large method-level shift; admitting M4c2 as a close competitor restores changes on 70 units without a comparable increase in mean independent-validator loss. A benchmark can therefore look decision-

stable while its numerical scores remain reference-dependent. This distinction is especially relevant when new models are compared against a mature baseline set: adding one dominant model may reduce ranking instability without removing the underlying score distortion.

Finally, the contrast results show why a single scalar correction factor is unlikely to be sufficient. T1, T2, and FLAIR preserve the direction of the effect but differ substantially in magnitude, and the five-method pool buffers T2 more strongly than T1. We therefore view reference dependence as a property to be measured for the operating regime of interest, not as a universal constant that can be subtracted from reported metrics. A practical benchmark can expose this sensitivity with a small repeated-acquisition subset even if most of the dataset has only one acquisition: the repeated subset can be used to estimate reference sensitivity, while the larger single-acquisition set supports conventional accuracy comparisons.

8 Discussion

The main conclusion is not that disjoint repetitions are clean ground truth—they remain noisy acquisitions—but that *independence and quality are different properties of a reference*. The all-slice and FLAIR extensions show that the selection effect is not an artifact of two hand-picked slices or one contrast. The learned-pool experiment adds two complementary boundaries: dominance can stabilize selection even when score shifts remain, and renewed competition can restore selection changes with modest independent-validator loss.

This interpretation complements the "hidden noise" result of Wang et al. (2024). Their comparison changes reference quality and processing, while ours holds repeat count and averaging fixed to isolate input overlap. It also connects naturally to self-supervised MRI methods that rely on statistically separated measurement partitions (Yaman et al., 2020). In all three cases, what is shared between input, target, and validation data determines the meaning of a reported risk estimate.

The external-validity limitation remains important. All measurements come from one 0.3T, four-coil dataset and one R=3 operating point. A bounded public-data search did not identify a second raw-k-space dataset with enough equal-status repeated acquisitions to reproduce the matched protocol. We therefore make no claim that the numerical effect sizes transfer to other scanners. The general claim is narrower: overlap is a measurable evaluation variable, and its consequence can be tested independently of reference averaging.

9 Limitations

The study uses one scanner/site and one acceleration operating point. FLAIR has only one independent validator because four repetitions are available. The learned extension contains one official PI-VarNet and one development-selected lightweight reconstruction, rather than a broad multi-site learned benchmark, and the 30-versus-24 ACS difference between its training setup and the evaluation mask is disclosed in the supplement. These constraints prevent a universal effect-size claim across scanners or architectures. They do not affect the matched-reference comparison within the reported M4Raw setting.

10 Conclusion

Reference overlap changes MRI reconstruction evaluation in two separable ways. It distorts the score of individual methods, including a modern learned reconstructor, and in competitive classical pools it can change which method is selected and make that choice worse on independent measurements. Full-volume T1/T2 and FLAIR results strengthen the empirical scope, while the contrasting one- and two-learned-candidate pools show why score distortion, winner changes, and held-out loss are separate endpoints. Benchmark reports should therefore treat reference provenance and independence as explicit parts of the evaluation design.

AI USE STATEMENT

Generative AI systems were used as research assistants for experiment orchestration, code review, literature organization, and language editing. All reported experimental values were produced by archived executable pipelines and were independently checked against frozen machine-readable records. The authors reviewed the scientific claims, statistical summaries, citations, and final manuscript and remain responsible for the work.

REPRODUCIBILITY STATEMENT

The appendix specifies the input mask, reconstruction parameters, participant partition, reference construction, validator construction, confidence-interval procedure, and controls. The supplementary artifact contains the scripts and machine-readable unit-level outputs needed to regenerate the paper tables and figures; no result in the main text is taken from an untracked manual calculation.

REFERENCES

Hemant K. Aggarwal, Merry P. Mani, and Mathews Jacob. Modl: Model-based deep learning architecture for inverse problems. *IEEE Transactions on Medical Imaging*, 38(2):394–405, 2019.

Vegard Antun, Francesco Renna, Clarice Poon, Ben Adcock, and Anders C. Hansen. On instabilities of deep learning in image reconstruction and the potential costs of ai. *Proceedings of the National Academy of Sciences*, 117(48):30088–30095, 2020.

Joshua Batson and Loic Royer. Noise2self: Blind denoising by self-supervision. In *International Conference on Machine Learning*, pp. 524–533, 2019.

Taejoon Eo, Yohan Jun, Taeseong Kim, Jinseong Jang, Ho-Joon Lee, and Dosik Hwang. Kiki-net: cross-domain convolutional neural networks for reconstructing undersampled magnetic resonance images. *Magnetic Resonance in Medicine*, 80(5):2188–2201, 2018.

Mark A. Griswold, Peter M. Jakob, Robin M. Heidemann, et al. Generalized autocalibrating partially parallel acquisitions (grappa). *Magnetic Resonance in Medicine*, 47(6):1202–1210, 2002.

Kerstin Hammernik, Teresa Klatzer, Erich Kobler, et al. Learning a variational network for reconstruction of accelerated mri data. *Magnetic Resonance in Medicine*, 79(6):3055–3071, 2018.

Thomas Janjusevic et al. Self-supervised noise-adaptive mri denoising via repetition to repetition learning. *Magnetic Resonance in Medicine*, 95(3):1619–1633, 2026.

Florian Knoll, Jure Zbontar, Anuroop Sriram, et al. fastmri: A publicly available raw k-space and dicom dataset of knee images for accelerated mr image reconstruction using machine learning. *Radiology: Artificial Intelligence*, 2(1):e190007, 2020.

Jaakko Lehtinen, Jacob Munkberg, Jon Hasselgren, et al. Noise2noise: Learning image restoration without clean data. In *International Conference on Machine Learning*, pp. 2965–2974, 2018.

Michael Lustig, David Donoho, and John M. Pauly. Sparse mri: The application of compressed sensing for rapid mr imaging. *Magnetic Resonance in Medicine*, 58(6):1182–1195, 2007.

Mengye Lyu, Lifeng Mei, Shoujin Huang, et al. M4raw: A multi-contrast, multi-repetition, multi-channel mri k-space dataset for low-field mri research. *Scientific Data*, 10:264, 2023.

Klaas P. Pruessmann, Markus Weiger, Markus B. Scheidegger, and Peter Boesiger. Sense: sensitivity encoding for fast mri. *Magnetic Resonance in Medicine*, 42(5):952–962, 1999.

Sathish Ramani, Thierry Blu, and Michael Unser. Monte-carlo sure: A black-box optimization of regularization parameters for general denoising algorithms. *IEEE Transactions on Image Processing*, 17(9):1540–1554, 2008.

Jo Schlemper, Jose Caballero, Joseph V. Hajnal, Anthony N. Price, and Daniel Rueckert. A deep cascade of convolutional neural networks for dynamic mr image reconstruction. *IEEE Transactions on Medical Imaging*, 37(2):491–503, 2018.

Elad Shimron, Jonathan I. Tamir, Ke Wang, and Michael Lustig. Implicit data crimes: Machine learning bias arising from misuse of public data. *Proceedings of the National Academy of Sciences*, 119(13), 2022.

Anuroop Sriram, Jure Zbontar, Tullie Murrell, et al. End-to-end variational networks for accelerated mri reconstruction. In *Medical Image Computing and Computer Assisted Intervention*, pp. 64–73, 2020.

Charles M. Stein. Estimation of the mean of a multivariate normal distribution. *The Annals of Statistics*, 9(6):1135–1151, 1981.

Martin Uecker, Peng Lai, Mark J. Murphy, et al. Espirit reconstruction in mri: Robust reconstruction with auto-calibrating parallel imaging. *Magnetic Resonance in Medicine*, 71(3):990–1001, 2014.

Jiayang Wang, Di An, and Justin P. Haldar. The "hidden noise" problem in mr image reconstruction. *Magnetic Resonance in Medicine*, 92(3):982–996, 2024.

Burhaneddin Yaman, Seyed Amir Hossein Hosseini, Steen Moeller, Jutta Ellermann, Kamil Ugurbil, and Mehmet Akcakaya. Self-supervised learning of physics-guided reconstruction neural networks without fully sampled reference data. *Magnetic Resonance in Medicine*, 84(6):3172–3191, 2020.

Jure Zbontar, Florian Knoll, Anuroop Sriram, et al. fastmri: An open dataset and benchmarks for accelerated mri. *arXiv preprint arXiv:1811.08839*, 2018.

Magauiya Zhussip, Shakarim Soltanayev, and Se Young Chun. Extending stein's unbiased risk estimator to train deep denoisers with correlated pairs of noisy images. In *Advances in Neural Information Processing Systems*, volume 32, 2019.