# WHAT DRIVES EARLY-STOPPING GAINS? THRESHOLD POLICIES AND AN MRI CASE STUDY

**Anonymous authors**
Paper under double-blind review

## ABSTRACT

Early-stopping systems use intermediate scores to decide when further computation or measurement is unnecessary. Yet greater resource savings can reflect better within-stage ranking, changes in cross-stage numerical scale, more flexible threshold policies, or a different fitting constraint, obscuring what improved. We study these dependencies for threshold-based stopping along fixed sequences, using repeated MRI as a case study. Sharing one threshold couples decisions across stages, so adding checkpoints can reduce attainable savings. Stage-specific thresholds can reproduce any shared-threshold policy that uses only a subset of checkpoints; with zero training failures, a greedy construction stops every training case no later than any feasible policy in this class. Strictly increasing stage-wise score transformations preserve stopping decisions when thresholds are transformed accordingly, distinguishing within-stage ranking from cross-stage numerical scale. Mechanism analyses in repeated prostate MRI show that these dependencies can change method comparisons: rescaling scores preserves within-stage ranking yet can reverse mean savings comparisons between scoring methods, while relaxing the training failure budget largely removes the advantage of fewer checkpoints. Brain MRI also exhibits scale sensitivity. The choice of baseline further changes the interpretation: in a separate test of prespecified policies, a strong fixed-budget control captures nearly all the primary adaptive policy's mean saving, leaving its additional advantage unresolved. Early-stopping savings thus reflect the score, stopping policy, and fitting procedure jointly; comparisons require explicit policy classes, matched fitting constraints, and strong fixed-budget controls.

## 1 INTRODUCTION

We consider early stopping in sequential computation or measurement, where a process terminates once an intermediate result is judged sufficient, avoiding unnecessary subsequent steps. Examples include early-exit neural networks that inspect predictions between layers and adaptive acquisition systems that inspect incoming measurements. Both combine a score with a rule that decides whether to stop. When one system saves more resources than another, which part improved? This paper studies how threshold rules affect that comparison, with repeated MRI providing a controlled case study.

Each position in a fixed sequence is a *stage*; a stage where stopping is allowed is a *checkpoint*. A score ranks cases at each stage, while a stopping *policy* specifies the checkpoints and numerical cutoffs. A *shared threshold* applies one cutoff across checkpoints; *stage-specific thresholds* allow a separate cutoff at each stage. Shared thresholds are used in practice: DeeBERT compares each layer's prediction entropy with one threshold, and FastBERT uses one uncertainty threshold across layers (Xin et al., 2020; Liu et al., 2020).

Sharing a threshold couples decisions across stages. When smaller scores favor stopping, avoiding an unacceptable early stop can force the cutoff low enough to reject otherwise acceptable later stops. Removing an early checkpoint may therefore increase savings without changing either the score or the observations. Stage-dependent score scaling creates a related dependency: multiplying all scores at one stage by the same positive factor preserves within-stage ranking but changes cross-stage numerical comparability, potentially changing savings under a shared threshold. Cross-stage



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.

numerical comparability can itself be useful; our interventions isolate scale sensitivity rather than assess probability calibration.

We organize the analysis around three controls. First, we characterize checkpoint restrictions through a counterexample and a policy-class inclusion result. Second, for stage-specific thresholds and zero training failures, we give an exact greedy construction that stops every training case no later than any feasible competitor in this class. Third, we establish decision preservation under strictly increasing stage-wise score transformations with correspondingly transformed thresholds. These results identify dependencies between score representation and policy choice; they do not yield an additive decomposition of gains.

Repeated MRI provides a setting in which these components can be changed while retaining the same observations. Measurements are repeated and averaged to reduce noise; fewer repetitions can shorten acquisition (Kaye et al., 2020). Our stopping policy uses only reconstructed images from the available repetitions, called the *prefix*. A stop is counted as a failure when its prefix violates a prespecified image-discrepancy criterion relative to the completed acquisition. The completed acquisition is used only retrospectively for fitting and evaluation. An initial study fits policies on 107 prostate patients and evaluates 12 frozen policies on 45 previously unused patients. A fixed eight-repeat control, fitted under a more permissive training failure budget, captures nearly all the primary adaptive policy's mean saving, showing how baseline choice changes the interpretation of adaptive gains. Later exploratory interventions on the 107 development patients and 25 previously accessed M4Raw brain subjects (Lyu et al., 2023) examine scale sensitivity and, for prostate MRI, matched training failure budgets. The theory concerns fixed score and loss trajectories; the interventions illustrate mechanisms, while the frozen test evaluates prespecified policies. Together, these results show that savings alone are insufficient to attribute improvement in threshold-based stopping. Comparisons should therefore use explicitly defined policy classes, matched fitting constraints, and strong fixed-budget controls.

## 2 Related work

### 2.1 Threshold-Based Early Stopping and Evaluation

DeeBERT and FastBERT use shared thresholds on uncertainty scores across layers to determine early exits (Xin et al., 2020; Liu et al., 2020). Under a shared threshold, changing the checkpoint set can change which stopping decisions the policy can represent. Comparisons across checkpoint sets therefore need to account for threshold flexibility.

Probability calibration concerns agreement between predicted confidence and observed correctness (Guo et al., 2017), whereas within-stage ranking and cross-stage numerical comparability describe different score properties. Fast yet Safe, Learn then Test, and risk-controlling prediction sets provide risk-control procedures under their respective assumptions (Jazbec et al., 2024; Angelopoulos et al., 2025; Bates et al., 2021). Neither our zero-training-failure optimality nor decision preservation under score transformations provides such a guarantee for unseen cases. Our focus is how scores, threshold policies, and fitting constraints interact when resource savings are used to compare stopping systems.

### 2.2 Adaptive MRI Acquisition

Adaptive MRI methods select encoding measurements or adapt sequence parameters (Zhang et al., 2019; Pineda et al., 2020; Bakker et al., 2020; Beracha et al., 2023), while other work jointly optimizes sampling and reconstruction (Oved & Shimron, 2026). Quality estimates and uncertainty also guide stopping or acquisition (Martinini et al., 2025; Fischer et al., 2025), while guided denoising reduces required diffusion-weighted MRI (DWI) averages (Kaye et al., 2020). Our case study holds the encoding sequence and reconstruction algorithm fixed. This lets us vary threshold rules and score scales on identical image trajectories, with a common reference and quality target.

## 3 Threshold-based stopping: setup and policy classes

We formalize threshold-based stopping along a fixed sequence to examine how checkpoint availability, threshold flexibility, and score representation across stages affect saving comparisons. We first define

the stopping policy and empirical fitting objective, then characterize the resulting policy classes and score transformations that preserve stopping decisions.

### 3.1 Decision rule and training objective

Consider cases $i$ observed along a fixed sequence of stages, with terminal stage $R$ and candidate checkpoints $\mathcal{K} \subseteq \{1, \dots, R-1\}$. At stage $k$, a real-valued score $z_i(k)$ uses only observations available through that stage; smaller values favor stopping. A binary loss $L_i(k) \in \{0, 1\}$ records whether stopping there fails the chosen quality requirement. Losses may depend on the completed trajectory and are available for fitting and evaluation, but not for online decisions. Stopping truncates the sequence without changing its preceding observations. The formal results hold the score and loss trajectories fixed.

A stopping policy selects a checkpoint set $\mathcal{C} \subseteq \mathcal{K}$ and thresholds $\boldsymbol{\tau} = (\tau_k)_{k \in \mathcal{C}}$. Its stopping time is

$$T_i(\mathcal{C}, \boldsymbol{\tau}) = \min\big(\{k \in \mathcal{C} : z_i(k) \le \tau_k\} \cup \{R\}\big). \tag{1}$$

The terminal stage is the fallback. A threshold of $-\infty$ disables a checkpoint; all others are finite. We compare *shared thresholds* $\tau_k = \tau$ on all candidate checkpoints or a selected subset, and *stage-specific thresholds* on all candidate checkpoints. Each cutoff in the last class can vary separately or disable its stage. We use *dense* for all candidate checkpoints and *sparse* for a proper subset.

Let $s_i(k)$ be the resource saving from stopping at stage $k$, nonincreasing in $k$, with $s_i(R) = 0$. On training cases $\mathcal{D}_{\mathrm{dev}}$, where $n_{\mathrm{dev}} = |\mathcal{D}_{\mathrm{dev}}|$, policy fitting within a specified class has objective

$$\max_{\mathcal{C}, \boldsymbol{\tau}} \frac{1}{n_{\mathrm{dev}}} \sum_{i \in \mathcal{D}_{\mathrm{dev}}} s_i(T_i) \quad \text{subject to} \quad \frac{1}{n_{\mathrm{dev}}} \sum_{i \in \mathcal{D}_{\mathrm{dev}}} L_i(T_i) \le \alpha_{\mathrm{dev}}. \tag{2}$$

The *training failure budget* $\alpha_{\mathrm{dev}}$ constrains the observed training failure fraction, not risk on new cases. Saving here depends only on the stopping stage; checkpoint-processing costs are handled separately in the MRI experiments.

### 3.2 Policy classes and score transformations

Checkpoint restrictions and threshold flexibility interact even with fixed scores. The following results characterize this interaction and the effect of changing score scale. The zero-failure training result assumes $L_i(R) = 0$ for every training case.

**Lemma 1** (Checkpoint availability and threshold flexibility)**.** *The class with all candidate checkpoints available and separate thresholds contains every shared-threshold rule on any checkpoint subset. The class using a shared threshold at all candidate checkpoints need not contain such a subset rule.*

Set $\tau_k = \tau$ on the chosen subset and $\tau_k = -\infty$ elsewhere to obtain the inclusion. For the second claim, consider checks at 4 and 7, with fallback at 12. Two cases have scores $(0.30, 0.15)$ and $(0.10, 0.08)$; both fail at 4 and pass at 7 and 12. Checking only at 7 with threshold 0.15 stops both without failure. With both checks enabled, any shared threshold at least 0.10 fails at 4; the best zero-failure rule can stop only the second case at 7. For a saving function with $s_i(7) = 0.30$ and $s_i(12) = 0$, mean saving falls from 30% to 15%. Disabling a checkpoint removes a stopping opportunity, not an observation, so checkpoint ablation under a shared threshold also changes the set of representable stopping decisions.

**Proposition 1** (Earliest feasible training stops)**.** *Fix finite scores and binary losses for each training case and stage, with zero loss at the full-budget fallback. At each candidate checkpoint in order, among cases not already stopped, let $u_k$ be the smallest score with $L_i(k) = 1$, or $+\infty$ if none exists. Accept every remaining case with score strictly below $u_k$, and choose the smallest threshold reproducing that set; disable the stage if it is empty. The resulting policy has zero training failures and stops every training case no later than any other zero-failure policy in the stage-specific threshold class.*

The proof uses induction on the remaining cases (App. A), without monotone scores or losses. The result implies optimal training saving for any reward nonincreasing in the stopping stage, but gives no such guarantee for checkpoint-dependent costs or unseen cases.

**Proposition 2** (Decision preservation under score transformation)**.** *Fix checkpoints and strictly increasing maps $g_k : \mathbb{R} \to \mathbb{R}$. Replacing $z_i(k)$ by $g_k(z_i(k))$ and each finite threshold $\tau_k$ by $g_k(\tau_k)$, while retaining disabled stages, preserves every stopping decision. Refitting Proposition 1 on the transformed training scores produces these mapped thresholds and hence the same decisions, including on unseen trajectories transformed by the same maps. A shared threshold generally cannot absorb different maps across stages.*

The equivalence $z_i(k) \le \tau_k \iff g_k(z_i(k)) \le g_k(\tau_k)$, together with the greedy tie convention, gives the result (App. A). This establishes decision preservation under matched transformations, not error control on new data; Section 6.1 uses it to isolate sensitivity to cross-stage scale while preserving within-stage ranking.

# 4 Repeated MRI case study

We instantiate the general stopping framework in repeated diffusion-weighted MRI, holding the acquisition order and reconstruction pipeline fixed. We define the MRI-specific loss and saving, then describe the scores and fitted policies used in the controlled comparisons.

## 4.1 Acquisition, quality target, and saving

In repeated DWI, a case is a patient and stage $k$ counts acquired high-$b$ repetitions per diffusion direction, where $b$ specifies diffusion weighting. For patient $i$, let $X_{d,r,s} \in \mathbb{R}^P$ denote the magnitude image for direction $d \in \{x, y, z\}$, repetition $r$, and slice $s$, restricted to a fixed region $\mathcal{B}$, with $S_i$ slices and $P$ pixels per region. We suppress patient indices on images and scores when unambiguous. Define $\bar{X}_{d,k,s} = k^{-1} \sum_{r=1}^{k} X_{d,r,s}$, with terminal budget $R = 12$. The reference $\bar{X}_{d,R,s}$ uses the same reconstruction pipeline as the prefix images. It is a noisy full-budget reference, not noise-free anatomy or a diagnostic ground truth.

The retrospective discrepancy and patient-level failure event are

$$\rho_{d,s}(k) = \frac{\|\bar{X}_{d,k,s} - \bar{X}_{d,R,s}\|_2}{\|\bar{X}_{d,R,s}\|_2 + \epsilon}, \qquad \epsilon = 10^{-8}, \tag{3}$$

$$L_i(k) = \mathbf{1}\left\{ \max_d \frac{1}{S_i} \sum_{s=1}^{S_i} \mathbf{1}\{\rho_{d,s}(k) > m\} > f \right\}. \tag{4}$$

Here $m$ is the relative discrepancy tolerance and $f$ the allowed fraction of slices exceeding it in each direction. The frozen prostate evaluation uses $m = 0.15$ and $f = 0.10$, fixed before test outcomes were evaluated. For an array $A = (A_{d,s})$, write $\mathcal{Q}_f(A) = \max_d A_{d,(q_i)}$, where $q_i = \lfloor f S_i \rfloor + 1$ and $A_{d,(q_i)}$ is the $q_i$-th largest value across slices. Then $L_i(k) = \mathbf{1}\{\mathcal{Q}_f(\rho(k)) > m\}$. A violation in *any* direction counts as a patient failure; pooling all $3S_i$ slices would weaken this event.

The primary region is a centered $50 \times 50$-pixel square (100 mm per side), an anatomical proxy rather than a prostate or lesion segmentation; the tolerance is an engineering choice without clinical validation. All slices enter Eq. 4. Secondary analyses evaluate whole-slice consistency and alternative tolerances at the same frozen stopping decisions.

The candidate checkpoints are $\mathcal{K} = \{4, \dots, 11\}$, with fallback $R = 12$. All four low-$b$ repetitions are available by stage 4. The released acquisition contains two $b = 0$ entries, four repetitions of three $b = 50$ directions, and twelve repetitions of three $b = 1000$ directions. Following the public reconstruction's array indexing, stage $k \ge 4$ consumes $14 + 3k$ of 50 entries. Under the header repetition time (TR), acquisition saving is

$$s_i^{\mathrm{acq}}(k) = \frac{50 - (14 + 3k)}{50} = 0.06(12 - k). \tag{5}$$

This saving covers the 50-entry DWI acquisition rather than the full MRI examination and excludes prescan and other fixed overhead; scanner event logs were unavailable to establish a prospective scheduler.

### 4.2 SCORES USED IN THE COMPARISONS

The MRI-specific scores $z_i(k)$ serve as inputs to the stopping policies. The primary residual score estimates the discrepancy between prefix and full-budget averages. For each direction and slice, define the repeat variance energy and residual score by

$$V_{d,s}(k) = \frac{1}{k-1}\sum_{r=1}^{k} \|X_{d,r,s} - \bar{X}_{d,k,s}\|_2^2, \qquad z_{\mathrm{res}}(k) = \mathcal{Q}_f\left(\frac{\sqrt{(R-k)V_{d,s}(k)/(Rk)}}{\|\bar{X}_{d,k,s}\|_2 + \epsilon}\right). \quad (6)$$

The factor $(R-k)/(Rk)$ follows from the covariance of the difference between overlapping averages under independent, identically distributed repetitions; App. B gives the derivation and Welford updates (Welford, 1962). The score uses only prefix images, with a threshold fitted from training losses; the model does not make $z_{\mathrm{res}} = m$ a validated quality cutoff.

Additional controls use adjacent-image change and its stage-dependent rescaling, a ridge predictor, and a Gaussian multiplier Monte Carlo score (Gaussian MC) based on a plug-in Gaussian model. Later mechanism analyses also use an Oracle Approximating Shrinkage (OAS)-based score (Chen et al., 2010). Definitions and fitting details appear in Apps. B and I. These scores provide alternative inputs to the stopping policies; the policy-class results do not depend on their construction.

### 4.3 FITTED POLICIES AND FIXED-BUDGET CONTROLS

MRI policy fitting uses $s_i(k) = s_i^{\mathrm{acq}}(k)$ in Eq. 2. The prespecified primary policy uses the residual score, a zero training failure budget, and a shared threshold at two checkpoints, $k = 7, 10$, selected on development data and frozen before test outcomes were evaluated. The original shared-threshold fits use a quantile grid; later mechanism analyses search all distinct training scores. App. B gives implementation details, while Proposition 1 provides the exact stage-specific construction.

Three fixed budgets provide controls: the full budget $R = 12$, the earliest fixed stage with zero training failures ($k = 10$), and the earliest fixed stage within a 5% training failure budget ($k = 8$, with 2/107 failures). The last control tests whether a fixed budget fitted more permissively already captures the adaptive saving; Section 6.2 compares policies under matched training failure budgets.

The remaining frozen policies comprise dense shared-threshold controls using adjacent, scaled adjacent, residual, and Gaussian MC scores; two-checkpoint shared-threshold controls using Gaussian MC and ridge; and two stage-specific residual controls using coordinate refinement and Proposition 1's greedy construction. App. C lists all 12 frozen policies and their parameters.

## 5 EXPERIMENTAL PROTOCOL

We distinguish development analyses, a frozen prostate test, and post-test mechanism analyses on prostate and brain MRI. We also specify the cost model and the assumptions underlying our uncertainty estimates.

**Data and roles.** The primary case study uses fastMRI prostate DWI (Tibrewala et al., 2024): 107 previously accessed patients constitute development data, including patients used in historical calibration or testing, and 45 previously unused patients form the frozen test. M4Raw brain MRI provides 25 historically accessed subjects solely for exploratory replication of scale sensitivity (Lyu et al., 2023).

We verified acquisition indexing, geometry, image validity, patient uniqueness, source integrity, and development–test separation, and performed an end-to-end reconstruction smoke test. The protocol was frozen before any new test outcomes were evaluated, and all 12 policies were evaluated before aggregate results were inspected. App. C records the technical exclusions and audit details.

**Reconstruction and causal scope.** Prostate images use the public GRAPPA/regridding and supplied-coil-map reconstruction (Griswold et al., 2002; Tibrewala et al., 2024), followed by averaging within each direction. No denoiser, registration, or reconstruction model is trained on test patients. Scores use only reconstructed prefixes; reference-dependent losses are evaluated after stopping decisions. The temporal provenance of the supplied coil sensitivity maps is unresolved: causality is enforced at the controller input, but end-to-end acquisition causality is not established.

**Development analyses.** Five-fold patient-level replay on the 107 development patients explores scores, checkpoint selection, and threshold flexibility, with all fitted quantities estimated within each training fold. Comparisons keep the reference, quality target, and acquisition denominator fixed. These analyses remain exploratory because the data informed the final design. App. D reports the original development controls; App. I reports subsequent score comparisons separately from the frozen evaluation.

**Post-test mechanism analyses.** After inspecting the original test results, we specified stage-scale interventions and matched training failure budgets on the 107 development patients. Three random five-fold partitions and a contiguous-ID partition assess sensitivity to fitting splits; repeated partitions do not increase the cohort size. M4Raw repeats the scale intervention separately on T1 and T2 images from the same 25 subjects, using six-repeat full-budget references and whole-slice root-sum-of-squares magnitude images. Each comparison fixes its reference and quality definition. These exploratory analyses report acquisition saving only and do not fit or evaluate new policies on the 45 test patients. App. J gives the protocols and complete results.

**Cost accounting.** For the frozen evaluation, the main cost model assumes acquisition pauses for measured image-domain updates, score computation, and executed threshold comparisons. For patient $i$, with total online cost $c_i$,

$$s_i^{\mathrm{img}} = s_i^{\mathrm{acq}}(T_i) - \frac{c_i}{50\,\mathrm{TR}_i},$$

where $c_i$ and $\mathrm{TR}_i$ are in seconds. Charges include feature updates through the last visited checkpoint, score computation, and executed threshold comparisons; fixed budgets have no online scoring cost. Each patient's execution is timed once, so these are implementation measurements rather than repeated hardware benchmarks. Raw reconstruction, disk I/O, scanner transfer, and command latency are excluded. Separate sensitivity analyses add per-check delays or one additional high-$b$ repetition after an adaptive stop request, without refitting policies. App. G specifies timing and delivery details.

**Endpoints and uncertainty.** The unit of analysis is the patient. Each frozen policy is evaluated by its failure count and a marginal one-sided 95% Clopper–Pearson upper bound $U_{.95}$ (Clopper & Pearson, 1934). The bounds assume independent Bernoulli patient outcomes with a common failure probability; they neither establish cohort representativeness nor provide simultaneous coverage across policies. Full-budget consistency loss is zero by construction.

Saving differences use 10,000 patient-paired bootstrap samples and percentile 95% intervals (App. E). These are approximate, separate contrasts, not risk-noninferiority or equivalence tests. Secondary quality assessment compares apparent diffusion coefficient (ADC) maps with the full-budget reference, computing each patient's mean signed error (bias), mean absolute error (MAE), and 95th percentile of absolute error across voxels. We report the across-patient mean of each metric; App. F defines ADC and the prefix-fixed mask.

# 6 Results

## 6.1 Shared-threshold comparisons are sensitive to stage scale

Changing cross-stage scale can alter shared-threshold savings without changing within-stage ranking. For each prostate score, we apply $z_i^{(\beta)}(k) = \exp(\beta(k - 7.5)/3.5)z_i(k)$, with $\beta \in \{-2, -1, 0, 1, 2\}$ fixed before this analysis. Every within-stage ranking and tie is preserved. We refit dense shared-threshold, two-checkpoint shared-threshold, and stage-specific policies on each training fold with zero training failures. Figure 1 averages three random partitions; ID-blocked results show the same qualitative pattern.

For the residual score, dense shared-threshold saving ranges from 0 to 18.22%, and two-checkpoint shared-threshold saving from 22.62 to 25.98%. Scale also reverses the observed comparison between scoring methods: under two-checkpoint shared thresholds, OAS exceeds Gaussian MC at $\beta = 0$ (27.72 versus 27.46%), but falls below it at $\beta = 1$ (26.07 versus 27.48%). Both have one out-of-fold failure per partition in these comparisons. This reversal demonstrates scale sensitivity; it neither establishes equal population risks nor attributes the original Gaussian MC advantage entirely to scale.

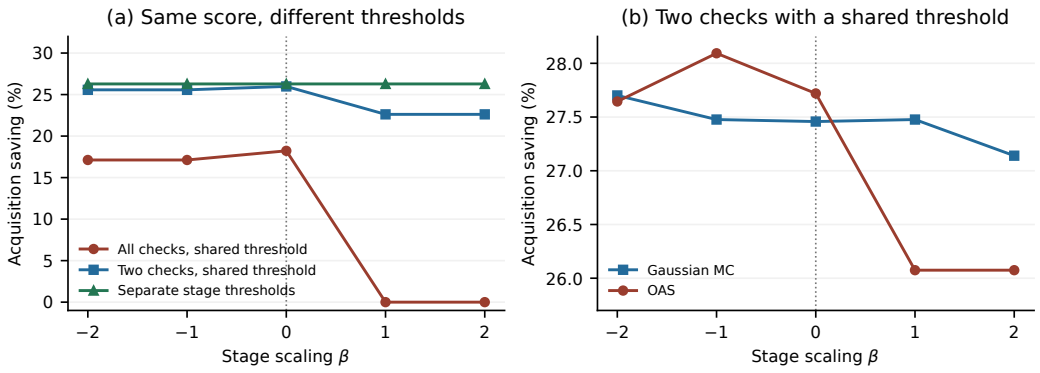



Figure 1: Score scaling changes savings while preserving within-stage ranking. (a) Residual score under three threshold policies. (b) Two scores under two-checkpoint shared thresholds, with a narrower vertical range. Exploratory analysis of 107 prostate patients; averages over three five-fold partitions. Savings exclude computation. Failures and complete comparisons: Table 9.

In contrast, stage-specific thresholds preserve every patient's stopping decision across the tested transformations, as predicted by Proposition 2; residual-score saving remains 26.28%. Decision preservation does not imply error control: stage-specific policies have 1–3 out-of-fold failures per partition across the tested scores and scales (Table 9).

The exploratory M4Raw analysis shows the same sensitivity in a second MRI setting. At tolerance 0.15, dense shared-threshold residual-score saving falls from 48.44% to 9.78% for T1 and from 53.56% to 14.22% for T2 when $\beta$ changes from 0 to 1. Stage-specific decisions remain unchanged. These percentages use six equal-cost repetitions; failure counts and alternative tolerances appear in App. J.1. Across both settings, shared-threshold comparisons change despite unchanged within-stage ranking.

### 6.2 Sparse-checkpoint gains depend on the training failure budget

The advantage of sparse checking largely disappears when the training failure budget is relaxed (Figure 2b,c). With zero training failures, residual-score dense and two-checkpoint policies save 18.22% and 25.98%, respectively, each with one out-of-fold failure per partition. At a 2.5% training failure budget, they save 29.51% and 29.38%, both with three failures per partition. The same pattern holds under ID-blocked validation. A fixed policy fitted under the same 2.5% budget saves 24% with two out-of-fold failures. Gaussian MC and OAS comparisons also lose their large sparse advantage. Thus the zero-training-failure result describes one fitting constraint, rather than a general advantage of fewer checkpoints; the additional saving at a relaxed budget must be read alongside the additional failures.

A separate hindsight diagnostic measures the opportunity for individual adaptation. Define $T_i^\star = \min\{k \in \mathcal{K} \cup \{R\} : L_i(k) = 0\}$, using each completed trajectory to identify its earliest acceptable stop. Across the 107 development patients, these stages yield 33.14% acquisition saving, whereas the earliest fixed stage with zero failures is $k = 10$ (Figure 2a). This is a ceiling for zero-failure saving on these observed trajectories, not a performance guarantee for an implementable policy. Approaching it requires recognizing acceptable prefixes without observing the remaining repetitions. Patient-specific stopping opportunities therefore coexist with strong dependence of the sparse-versus-dense comparison on the training failure budget.

### 6.3 Frozen evaluation: the role of fixed-budget controls

A stronger fixed-budget control captures nearly all the primary adaptive policy's mean saving in the frozen test. The primary policy stops 31 patients at $k = 7$ and 14 at $k = 10$, saving 24.40% of acquisition and 24.39% after charging image-domain computation (Table 1). Against the zero-training-failure fixed $k = 10$ policy, its advantage is 12.39 percentage points, with paired 95% interval

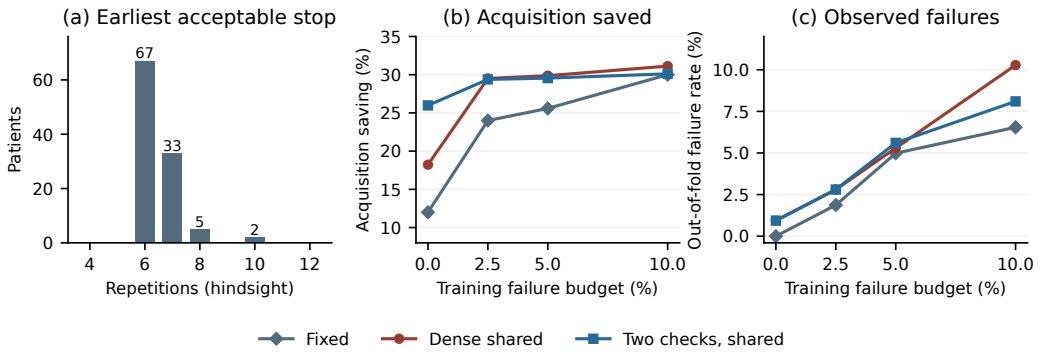



Figure 2: Individual stopping opportunities and matched training failure budgets ($n = 107$). (a) Earliest acceptable stage using the completed acquisition. (b,c) Savings and failures for residual-score and fixed policies, averaged over three five-fold partitions. Training budgets constrain fitting; observed failures are evaluated out of fold. Full results: Table 10.

Table 1: Frozen-policy evaluation on 45 test patients. Acq.: saving under the 50-entry budget. Image: saving after measured image-domain computation. $U_{.95}$: marginal 95% risk upper bound. Bold identifies the prespecified primary policy.

| Frozen policy | Acq. (%) | Image (%) | Failures | $U_{.95}$ (%) | Cost (s) |
|---|---|---|---|---|---|
| Full budget | 0.00 | 0.00 | 0/45 | 0 (struct.) | 0.000 |
| Fixed $k = 10$ (zero training failures) | 12.00 | 12.00 | 0/45 | 6.44 | 0.000 |
| Fixed $k = 8$ (5% training budget) | 24.00 | 24.00 | 0/45 | 6.44 | 0.000 |
| Adjacent, shared dense | 13.33 | 13.33 | 0/45 | 6.44 | 0.022 |
| Scaled adjacent, shared dense | 13.87 | 13.86 | 0/45 | 6.44 | 0.021 |
| Residual, shared dense | 17.20 | 17.19 | 0/45 | 6.44 | 0.020 |
| **Residual, two checks (primary)** | **24.40** | **24.39** | **0/45** | **6.44** | **0.016** |
| Residual, coordinate thresholds | 25.73 | 25.73 | 0/45 | 6.44 | 0.015 |
| Residual, greedy stage thresholds | 25.73 | 25.73 | 0/45 | 6.44 | 0.015 |
| Ridge, two checks | 24.00 | 23.99 | 0/45 | 6.44 | 0.016 |
| Gaussian MC, shared dense | 17.20 | 14.45 | 0/45 | 6.44 | 8.143 |
| Gaussian MC, two checks | 26.80 | 26.20 | 0/45 | 6.44 | 1.791 |

$[9.99, 14.80]$. Against fixed $k = 8$, which saves 24%, the difference falls to 0.39 points, with interval $[-2.01, 2.80]$. This interval leaves the additional adaptive advantage unresolved.

The primary and fixed $k = 8$ policies both have zero test failures, but were fitted with different training failure budgets. Their observed failure counts do not establish equal population risks. For the primary policy, zero failures among 45 patients gives a one-sided 95% risk upper bound of 6.44%, insufficient to establish a 5% population-risk guarantee.

**Checkpoint and threshold controls.** The primary policy's gain over dense residual checking can be recovered within the stage-specific threshold class, which allows all candidate checkpoints. The primary exceeds the dense shared-threshold policy by 7.20 points, with paired interval $[4.93, 9.20]$. Both stage-specific controls attain 25.73% image-domain saving, 1.33 points above the primary; the primary-minus-greedy interval is $[-2.13, -0.67]$. Development controls show the same pattern (App. D). Thus threshold flexibility must be accounted for in the sparse-versus-dense comparison; these results do not assign additive contributions to policy components.

**Score and computation controls.** Gaussian MC illustrates how checkpoint choice affects both acquisition saving and computation cost. Moving from dense shared thresholds to two selected checkpoints raises acquisition saving from 17.20% to 26.80% and image-domain saving from 14.45% to 26.20%, while mean online computation cost falls from 8.14 s to 1.79 s.

Within the two-checkpoint shared-threshold class, Gaussian MC exceeds the primary residual policy by 1.80 points in image-domain saving; the primary-minus-Gaussian MC interval is $[-3.80, -0.16]$.

The residual policy costs 15.52 ms per patient on average, while ridge reaches 23.99% image-domain saving, with no clear gain over the residual policy. Further covariance-score comparisons and temporal model stress tests appear in App. I. The observed Gaussian MC advantage is conditional on this policy class and cost model; both within-stage ranking and cross-stage numerical scale can contribute.

### 6.4 Quality and delivery sensitivity

Sensitivity analyses retain the frozen policies and examine how their results depend on the quality target and delivery model. For the primary policy, the across-patient means of ADC bias, MAE, and within-patient 95th-percentile absolute error are 2.47, 26.20, and 66.50 in units of $10^{-6}$ mm$^2$/s. Fixed $k = 8$ has a slightly lower mean ADC MAE of $25.72 \times 10^{-6}$ mm$^2$/s. Figure 4 shows cases selected by a fixed descriptive rule after testing.

At the stricter consistency tolerance $m = 0.10$, unchanged primary decisions fail in 31/45 patients, versus 42/45 for fixed $k = 8$. At $m = 0.20$, all frozen policies have zero failures; whole-slice evaluation at $m = 0.15$ records one primary failure. The nominal zero-failure result is therefore specific to the prespecified regional quality target.

An additional one-second transfer delay per executed check reduces primary saving to 23.95%, below fixed $k = 8$. Delivering one additional high-$b$ repetition after an adaptive stop request reduces its acquisition-only saving to 18.40%, while fixed $k = 8$ remains at 24%. Every policy still has zero primary-endpoint failures on the delivered images in this scenario. These analyses leave the original frozen comparison unchanged and expose its dependence on delivery assumptions. A controller trained for delayed delivery would solve a different optimization problem; App. G reports all cost scenarios.

## 7 Discussion and conclusion

Early-stopping savings are joint properties of the score, stopping policy, and fitting procedure. Shared thresholds couple decisions across stages, allowing checkpoint availability, cross-stage numerical scale, and threshold flexibility to change savings even when within-stage ranking is unchanged. Fitting constraints further alter these comparisons. The formal results characterize these dependencies, and the MRI interventions show that they can materially change empirical conclusions. Cross-stage numerical comparability can itself be useful; its effect must be distinguished from improved within-stage ranking.

The frozen evaluation demonstrates the practical consequence: a strong fixed-budget control captures nearly all the primary adaptive policy's mean saving, materially changing the interpretation of its gain over a more conservative baseline. This finding does not establish policy equivalence: the paired interval spans both signs, and the policies were fitted under different training failure budgets. Likewise, the gain over dense shared-threshold checking is recovered within the stage-specific threshold class, which allows all candidate checkpoints. Together, these results show why an observed gain cannot be attributed to a better score or to adaptation itself without accounting for policy design, fitting constraints, and baseline choice.

**Scope and limitations.** The MRI endpoint measures agreement with a noisy full-budget reference using a proxy region and an engineering tolerance, without clinical validation, lesion annotations, or reader assessment. The frozen test comprises 45 patients from contiguous blocks of one public prostate source; subsequent prostate and brain analyses use historical data and remain exploratory. End-to-end acquisition causality, reconstruction and transfer costs, and scanner delivery require prospective verification, and empirical effects outside MRI remain untested. The formal results characterize policy classes and decision preservation under their stated assumptions; they provide no population-risk guarantee.

Meaningful comparisons of threshold-based stopping require explicit policy classes, matched fitting constraints, and strong fixed-budget controls, with savings interpreted jointly with failures and computation costs. Without these controls, larger resource savings alone do not establish a better score or more effective adaptation.

### AI USE STATEMENT

We used a generative AI assistant for literature search, research planning and execution, and manuscript drafting and language editing. Research assistance included feedback on experimental design, software implementation, analysis of results, and review of technical arguments. Verification included reference checks, software tests, and comparisons with experimental outputs. The authors take responsibility for the final manuscript and all reported results.

### ETHICS STATEMENT

This is a retrospective analysis of de-identified fastMRI prostate and public M4Raw brain data, with no new participant recruitment or data collection. Data access and reuse are subject to the respective dataset terms. The central prostate region is an anatomical proxy, not a clinical annotation. The policies have not been deployed prospectively or validated for diagnostic decisions; clinical use requires appropriate endpoints, reader assessment, and prospective validation.

### REPRODUCIBILITY STATEMENT

App. A states the theoretical assumptions and proofs; App. B specifies reconstruction, scores, and policy fitting. App. C records frozen parameters, cohort membership, technical exclusions, timestamps, and integrity checks. The frozen evaluation records code, environment, input hashes, and policy parameters before any new test outcomes were evaluated; its audit includes identical-protocol resumption checks. Apps. D–G document development controls, paired comparisons, secondary quality metrics, and timing. The later exploratory analyses on historically accessed data have separate protocols and results in Apps. I and J.

## REFERENCES

Anastasios N. Angelopoulos, Stephen Bates, Emmanuel J. Candès, Michael I. Jordan, and Lihua Lei. Learn then test: Calibrating predictive algorithms to achieve risk control. *The Annals of Applied Statistics*, 19(2):1641–1662, 2025. doi: 10.1214/24-AOAS1998.

Tim Bakker, Herke van Hoof, and Max Welling. Experimental design for MRI by greedy policy search. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 33, 2020.

Stephen Bates, Anastasios Angelopoulos, Lihua Lei, Jitendra Malik, and Michael I. Jordan. Distribution-free, risk-controlling prediction sets. *Journal of the ACM*, 68(6):43:1–43:34, 2021. doi: 10.1145/3478535.

Inbal Beracha, Amir Seginer, and Assaf Tal. Adaptive model-based magnetic resonance. *Magnetic Resonance in Medicine*, 90(3):839–851, 2023. doi: 10.1002/mrm.29688.

Yilun Chen, Ami Wiesel, Yonina C. Eldar, and Alfred O. Hero, III. Shrinkage algorithms for MMSE covariance estimation. *IEEE Transactions on Signal Processing*, 58(10):5016–5029, 2010. doi: 10.1109/TSP.2010.2053029.

Charles 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.

Paul Fischer, Jan Nikolas Morshuis, Thomas Küstner, and Christian Baumgartner. CUTE-MRI: Conformalized uncertainty-based framework for time-adaptive MRI. *arXiv preprint arXiv:2508.14952*, 2025.

Mark A. Griswold, Peter M. Jakob, Robin M. Heidemann, Mathias Nittka, Vladimir Jellus, Jianmin Wang, Berthold Kiefer, and Axel Haase. Generalized autocalibrating partially parallel acquisitions (GRAPPA). *Magnetic Resonance in Medicine*, 47(6):1202–1210, 2002. doi: 10.1002/mrm.10171.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. On calibration of modern neural networks. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pp. 1321–1330, 2017. URL https://proceedings.mlr.press/v70/guo17a.html.

Metod Jazbec, Alexander Timans, Tin Hadži Veljković, Kaspar Sakmann, Dan Zhang, Christian A. Naesseth, and Eric Nalisnick. Fast yet safe: Early-exiting with risk control. In *Advances in Neural Information Processing Systems*, volume 37, 2024. doi: 10.52202/079017-4124.

Elena A. Kaye, Emily A. Aherne, Cihan Duzgol, Ida Häggström, Erich Kobler, Yousef Mazaheri, Maggie M. Fung, Zhigang Zhang, Ricardo Otazo, Hebert A. Vargas, and Oguz Akin. Accelerating prostate diffusion-weighted MRI using a guided denoising convolutional neural network: Retrospective feasibility study. *Radiology: Artificial Intelligence*, 2(5):e200007, 2020. doi: 10.1148/ryai.2020200007.

Weijie Liu, Peng Zhou, Zhiruo Wang, Zhe Zhao, Haotang Deng, and Qi Ju. FastBERT: a self-distilling BERT with adaptive inference time. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 6035–6044, 2020. doi: 10.18653/v1/2020.acl-main.537.

Mengye Lyu, Lifeng Mei, Shoujin Huang, Sixing Liu, Yi Li, Kexin Yang, Yilong Liu, Yu Dong, Linzheng Dong, and Ed X. Wu. M4Raw: A multi-contrast, multi-repetition, multi-channel MRI k-space dataset for low-field MRI research. *Scientific Data*, 10:264, 2023. doi: 10.1038/s41597-023-02181-4.

Filippo Martinini, Mauro Mangia, Alex Marchioni, Gianluca Setti, and Riccardo Rovatti. Incremental undersampling MRI acquisition with neural self assessment. *Signal Processing*, 228:109746, 2025. doi: 10.1016/j.sigpro.2024.109746.

Tal Oved and Efrat Shimron. NexOP: Joint optimization of NEX-aware k-space sampling and image reconstruction for low-field MRI. *arXiv preprint arXiv:2605.11583*, 2026.

Luis Pineda, Sumana Basu, Adriana Romero, Roberto Calandra, and Michal Drozdzal. Active MR k-space sampling with reinforcement learning. In *Medical Image Computing and Computer Assisted Intervention (MICCAI)*, volume 12262 of *Lecture Notes in Computer Science*, pp. 23–33, 2020. doi: 10.1007/978-3-030-59713-9_3.

Radhika Tibrewala, Tarun Dutt, Angela Tong, Luke Ginocchio, Riccardo Lattanzi, Mahesh B. Keerthivasan, Steven H. Baete, Sumit Chopra, Yvonne W. Lui, Daniel K. Sodickson, Hersh Chandarana, and Patricia M. Johnson. FastMRI Prostate: A public, biparametric MRI dataset to advance machine learning for prostate cancer imaging. *Scientific Data*, 11:404, 2024. doi: 10.1038/s41597-024-03252-w.

B. P. Welford. Note on a method for calculating corrected sums of squares and products. *Technometrics*, 4(3):419–420, 1962. doi: 10.1080/00401706.1962.10490022.

Ji Xin, Raphael Tang, Jaejun Lee, Yaoliang Yu, and Jimmy Lin. DeeBERT: Dynamic early exiting for accelerating BERT inference. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 2246–2251, 2020. doi: 10.18653/v1/2020.acl-main.204.

Zizhao Zhang, Adriana Romero, Matthew J. Muckley, Pascal Vincent, Lin Yang, and Michal Drozdzal. Reducing uncertainty in undersampled MRI reconstruction with active acquisition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2049–2058, 2019.

# A POLICY-CLASS ARGUMENTS

We use the fixed score and loss trajectories of Section 3. The arguments concern representable stopping decisions and empirical fitting; none assumes a probabilistic model for the scores.

**Details for Lemma 1.** For any sparse shared-threshold rule $(\mathcal{C}, \tau)$, set $\tau_k = \tau$ on $\mathcal{C}$ and $\tau_k = -\infty$ elsewhere. This represents the same stopping decisions in the stage-specific threshold class with all candidate checkpoints available. No distributional or monotonicity assumptions on the score and loss trajectories are needed. The corresponding inclusion need not hold for the dense shared-threshold class, as the two-case example in Section 3.2 demonstrates. For that example, thresholds in $[0.08, 0.10)$ stop only the second case at 7, while thresholds below 0.08 stop neither; every threshold at least 0.10 violates zero training failure at stage 4. The example holds the fallback $R$ and saving function fixed.

**Proof of Proposition 1.** Let $\mathcal{A}_k$ be the cases not stopped before stage $k$ by the greedy construction, and $\mathcal{A}_k^\pi$ the analogous set for any feasible competing policy $\pi$. Initially the sets agree. Assume $\mathcal{A}_k \subseteq \mathcal{A}_k^\pi$. A greedy acceptance has score below every score with $L_i(k) = 1$ in $\mathcal{A}_k$, so its loss is zero. If a competing threshold accepts case $i \in \mathcal{A}_k$ at this stage, it cannot reach the score of any case with $L_i(k) = 1$ in $\mathcal{A}_k^\pi$. It must therefore be below every such score in $\mathcal{A}_k$. Thus $i$ is also accepted by the greedy construction. Cases already removed from $\mathcal{A}_k$ have stopped even earlier. The remaining-set inclusion therefore holds at the next checkpoint. Induction gives zero greedy failures and no-later stopping for every case. When several thresholds reproduce an acceptance set, taking the largest accepted observed score is the smallest threshold that includes that set; ties with a failing case's score must be rejected. An empty set is represented by $-\infty$. The fallback has zero loss by assumption. □

The conclusion holds for the given finite training sample and scores at zero training failure budget. The result does not establish optimality for positive training failure budgets, checkpoint-dependent costs, or unseen cases. Choosing the smallest threshold for each acceptance set fixes otherwise equivalent training solutions. The implementation is checked against exhaustive threshold enumeration on 40 small synthetic problems.

**Proof of Proposition 2.** Strict increase preserves each comparison with a finite threshold, and a disabled stage remains disabled. Taking the first accepted stage therefore preserves stopping on every trajectory. For the refitted greedy rule, assume both runs have the same remaining training cases at stage $k$. Order and ties are identical after transformation, so the failing cases define the same strict acceptance set. Its smallest threshold is the largest accepted score, which transforms to $g_k(\tau_k)$; an empty set disables the stage in both runs. Induction gives the mapped thresholds at every stage. They preserve decisions on new trajectories as well, provided the same $g_k$ are applied. This establishes decision preservation, not a bound on new losses. It does not establish refitting invariance for arbitrary fitting algorithms, finite quantile grids, or regularized threshold objectives. For shared thresholds, stage-dependent maps can produce unequal cutoffs; the common-threshold restriction prevents absorbing them in general. For example, in Lemma 1, multiply stage-4 scores by 10 and leave stage 7 unchanged. Within-stage rankings stay fixed, but a shared threshold 0.15 now stops both cases at 7 without failure. □

# B IMPLEMENTATION DETAILS

This appendix specifies the reconstruction, scores, and fitting procedures used in the frozen prostate evaluation. Later exploratory score controls and exact shared-threshold searches are described in Apps. I and J.

**Reconstruction.** The public implementation performs regridding, GRAPPA interpolation, coil combination using the supplied sensitivity maps, and magnitude reconstruction. Our per-average arrays have shape $(50, S_i, 200, 150)$; the primary region $\mathcal{B}$ is rows 75:125 and columns 50:100 in zero-based, half-open indexing. We verify the acquisition table against the release's averaging indices. The controller's region is fixed in this array coordinate system. The smoke test on raw development patient 001 matches the previously derived per-average images and independently invokes the public reconstruction on its first slice. Six direction/$b$ averages agree to relative errors

$7.94 \times 10^{-9}$–$1.35 \times 10^{-8}$. This validates the implemented reconstruction path against that code, not equivalence to a scanner vendor's stored reconstruction or the temporal origin of its sensitivity maps.

**Residual-score model.** Holding patient, direction, and slice fixed, suppose repetitions are independent and identically distributed with finite covariance $\Sigma$. Then

$$\mathrm{Cov}(\bar{X}_k - \bar{X}_R) = \frac{R-k}{Rk}\Sigma, \qquad \mathrm{Cov}(\bar{X}_k - \bar{X}_{k-1}) = \frac{\Sigma}{k(k-1)}. \tag{7}$$

The first identity accounts for the observations shared by the prefix and full-budget averages. Substituting the observed variance energy motivates Eq. 6. Magnitude noise and changing acquisition conditions can violate the model, and prefix normalization replaces the unavailable reference norm. Thresholds are therefore fitted from training outcomes.

**Derivation of Eq. 7.** For $k < R$ and independent repetitions with common covariance $\Sigma$, write $\bar{X}_R = (k/R)\bar{X}_k + ((R-k)/R)\bar{X}_{\mathrm{suffix}}$. The prefix and suffix averages are independent, hence

$$\mathrm{Cov}(\bar{X}_k - \bar{X}_R) = \left(\frac{R-k}{R}\right)^2 \left(\frac{\Sigma}{k} + \frac{\Sigma}{R-k}\right) = \frac{R-k}{Rk}\Sigma.$$

Likewise, $\bar{X}_k - \bar{X}_{k-1} = (X_k - \bar{X}_{k-1})/k$ has covariance $\Sigma/[k(k-1)]$. These are unconditional covariance identities. Replacing the true covariance with observed residuals and conditioning a future-image simulator at $\bar{X}_k$ are additional modeling steps.

**Streaming updates.** Suppress direction and slice indices. Initialize $\bar{X}_1 = X_1$ and $M_{2,1} = 0$. For each incoming image $X_k$, $k \geq 2$, set $\Delta_k = X_k - \bar{X}_{k-1}$ and $\bar{X}_k = \bar{X}_{k-1} + \Delta_k/k$. The per-slice residual-energy accumulator updates by $M_{2,k} = M_{2,k-1} + \langle \Delta_k, X_k - \bar{X}_k \rangle$. Then $V(k) = M_{2,k}/(k-1)$. The squared adjacent-change norm is $\|\Delta_k/k\|_2^2$. The implementation computes the complete eight-feature vector at visited feature checkpoints, even when a policy uses only one component; this cost is included.

**Adjacent-change controls.** The adjacent score is $\mathcal{Q}_f(\|\bar{X}_k - \bar{X}_{k-1}\|_2/(\|\bar{X}_k\|_2 + \epsilon))$. The scaled adjacent score multiplies it by $\sqrt{(k-1)(R-k)/R}$, as suggested by Eq. 7. This covariance correction need not make the nonlinear order-statistic scores equivalent. These controls, the ridge predictor, and Gaussian MC all use the same observed prefixes.

**Ridge features and fit.** The eight features are $k/R$, $1/k$, adjacent score, scaled adjacent score, residual score, median per-slice residual prediction, direction-maximum slice quantile of relative repeat standard deviation, and the range across directions of the residual quantiles. For stages 4–11, features are clipped below at $10^{-8}$ and logged, standardized using training means and standard deviations (the latter floored at $10^{-6}$), and fitted to $\log\max\{\mathcal{Q}_f(\rho(k)), 10^{-8}\}$ with ridge penalty 1. The intercept is the training target mean; the fitted log prediction is clipped to $[-30, 5]$ before exponentiating. Cross-validation estimates all fitted quantities within the training fold; the frozen fit uses all 107 development patients. The ridge penalty is fixed before test outcomes are evaluated. Its policy uses the same two-checkpoint search.

**Gaussian multiplier Monte Carlo.** At a fixed stage $k$, define residuals $E^{(k)}_{d,r,s} = X_{d,r,s} - \bar{X}_{d,k,s}$ and draws

$$\widetilde{X}^{(j)}_{d,R,s} = \bar{X}_{d,k,s} + \sqrt{\frac{R-k}{Rk(k-1)}} \sum_{r=1}^{k} W^{(j)}_r E^{(k)}_{d,r,s}, \qquad W^{(j)}_r \sim N(0,1).$$

Multipliers are independent over $r$ and $j$; each draw is shared across all directions and slices. The score averages, over $B = 200{,}000$ draws, the indicator that $\mathcal{Q}_f(\|\widetilde{X}^{(j)}_R - \bar{X}_k\|_2/(\|\widetilde{X}^{(j)}_R\|_2 + \epsilon)) > m$. Here $j$ indexes draws. The base seed 20260921 and stage-dependent seed derivation are fixed before test outcomes are evaluated and shared across patients. Sufficient statistics consist of the current mean norm, its residual inner products, and residual Gram matrices per slice, allowing the Gaussian draws to be evaluated without materializing $B$ volumes. The implementation uses chunks of 4096 draws. The draws define a deterministic implemented score once seeded; they are not independent clinical replicates.

**Threshold grids and ties.** The original frozen dense and sparse shared-threshold fits use 129 pooled quantiles over stages 4–11, plus $-\infty$. Sparse search enumerates lexicographically ordered checkpoint combinations and ascending unique thresholds. The first maximizer of training saving among feasible candidates is retained. Coordinate refinement starts from the primary sparse vector, makes at most two passes over stages 4–11, and considers $-\infty$, the current threshold, and 33 per-stage quantiles. It accepts strictly improved feasible training saving, with tolerance $10^{-12}$ in the objective comparison. This coordinate procedure gives a local refinement. The exact greedy stage-specific control uses Proposition 1, including its largest-accepted-score tie convention. Later exploratory searches enumerate all distinct training scores rather than the frozen quantile grid.

## C FROZEN PARAMETERS AND TECHNICAL EXCLUSIONS

The parameters below were fitted on development data and frozen before the 45 test patients' outcomes were evaluated. Both stage-specific policies allow all candidate checkpoints during fitting; their fitted finite thresholds occur only at $k = 6, 7, 9, 10$. Other stages are disabled.

Table 2: Frozen parameters fitted on 107 development patients. Values are rounded; the saved protocol retains full precision. Ridge thresholds apply to the fitted predictor.

| Frozen policy | Stage: threshold (disabled stages omitted) |
|---|---|
| Full budget | fixed $k = 12$ |
| Fixed $k = 10$ (zero training failures) | fixed $k = 10$ |
| Fixed $k = 8$ (5% training budget) | fixed $k = 8$ |
| Adjacent, shared dense | $k = 4, \ldots, 11$: 0.0583749 |
| Scaled adjacent, shared dense | $k = 4, \ldots, 11$: 0.0823717 |
| Residual, shared dense | $k = 4, \ldots, 11$: 0.084759 |
| Residual, two checks (primary) | 7: 0.12251, 10: 0.12251 |
| Residual, coordinate thresholds | 6: 0.131986, 7: 0.122515, 9: 0.0848587, 10: 0.12251 |
| Residual, greedy stage thresholds | 6: 0.134828, 7: 0.122511, 9: 0.0849563, 10: 0.0839797 |
| Ridge, two checks | 7: 0.128348, 10: 0.128348 |
| Gaussian MC, shared dense | $k = 4, \ldots, 11$: 0.00248867 |
| Gaussian MC, two checks | 7: 0.186683, 10: 0.186683 |

The 107 complete development patients come from the previously accessed ID range 001–132. The 45 retained test patients have 26–36 slices, 2 mm in-plane spacing, and header TRs of 5.0–6.9 seconds. The five test archive blocks contribute 10, 9, 8, 8, and 10 patients, respectively, in the ranges 133–142, 143–152, 153–162, 163–172, and 173–182. IDs 147, 158, and 166 occur in incomplete members; 160 and 164 are absent from the readable portions. No patient was excluded based on stopping performance, image error, or ADC error. Truncation is distinguished from a compressed-stream checksum failure: a CRC mismatch would require separate integrity resolution and is not accepted as a simple missing tail. There are no unresolved reconstruction or derivation errors in the retained cohort. Complete images are checked for correct 50-entry indexing, finite nonnegative values, nonempty signal, required geometry, matching header TR, and unique raw-member hashes. The end-to-end smoke test reconstructs a development patient from raw data, compares against the public implementation, evaluates all 12 policies on development patients with different slice counts, and verifies identical resumed results.

The protocol's SHA-256 is

`23dbd72745c5109ca3897ea832d824648a23c16dd56c78d41e0377a14eb566ba`

The protocol was frozen at 2026-09-22 01:11:12 UTC. Test-outcome evaluation began at 01:12:39 UTC and finished at 01:23:29 UTC on the same date. The protocol includes all 12 policies, the 45 retained patients, five technical exclusions, development input hashes, code hashes, environment, and completed audit and smoke-test records. Per-patient outputs carry the same protocol hash. A final manifest binds all result JSON files. These local integrity records are useful for reproducibility; statistical independence also relies on the stated research history and cannot be inferred from hashes alone.

## D ORIGINAL EXPLORATORY DEVELOPMENT CONTROLS

These controls use five-fold patient-level replay on the 107 development patients to examine checkpoint choice, threshold flexibility, and score construction under a zero training failure budget. All saving percentages exclude computation.

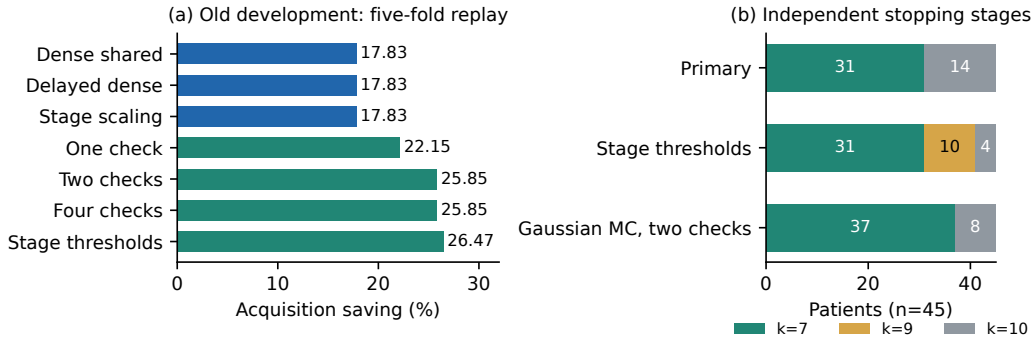



Figure 3: Threshold flexibility and stopping stages. (a) Residual-score cross-validation on 107 development patients; each policy has zero training failures and one out-of-fold failure. (b) Stopping stages of three frozen policies on 45 test patients.

Table 3: Exploratory five-fold development comparison ($n = 107$), fitted with zero training failures. Quality uses the central region at $m = .15$; savings use Eq. 5 and exclude computation.

| Exploratory five-fold policy | Acquisition saving (%) | Failures / 107 |
|---|---|---|
| Adjacent, dense | 16.15 | 2 |
| Scaled adjacent, dense | 16.49 | 2 |
| Residual, dense | 17.83 | 1 |
| Residual, delayed dense | 17.83 | 1 |
| Residual, median stage scaling | 17.83 | 1 |
| Residual, upper stage scaling | 17.83 | 1 |
| Residual, selected one check | 22.15 | 1 |
| Residual, selected two checks | 25.85 | 1 |
| Residual, selected four checks | 25.85 | 1 |
| Residual, coordinate thresholds | 26.52 | 1 |
| Residual, greedy stage thresholds | 26.47 | 1 |
| Ridge, selected two checks | 26.47 | 2 |
| Gaussian MC, dense | 18.90 | 1 |
| Gaussian MC, selected two checks | 27.03 | 1 |

The fixed zero-training-failure policy saves 12.00% with 0/107 out-of-fold failures. The residual delayed-dense control chooses the first allowed stage on the training fold and then checks every remaining stage. Stage scaling multiplies each stage's score by a training quantile of realized discrepancy divided by score, using the median or 90th percentile. The lack of improvement from these two simple corrections is an empirical finding, not proof that no stage-specific rescaling can work. The stage-specific threshold results show that a class allowing all candidate checkpoints can recover the saving advantage of sparse shared-threshold checking; the fitted policy may still disable stages.

The two-checkpoint residual selection uses $(7, 10)$ in four folds and $(7, 8)$ in one fold. All policies selected for the frozen test are refitted on the full 107-patient development set before test outcomes are evaluated. The two-checkpoint residual policy retains its prespecified primary status. Earlier tolerance and training failure budget sweeps remain exploratory; they do not change the frozen quality target.

## E PAIRED COMPARISONS ON THE FROZEN TEST

Table 4: Primary-minus-baseline saving differences in percentage points ($n = 45$). Marginal 95% intervals use 10,000 patient-paired percentile bootstrap samples. Savings include measured image-domain computation.

| Baseline | Primary minus baseline (pp) | Paired 95% interval |
|---|---|---|
| Full budget | 24.39 | [21.99, 26.80] |
| Fixed $k = 10$ (zero training failures) | 12.39 | [9.99, 14.80] |
| Fixed $k = 8$ (5% training budget) | 0.39 | [-2.01, 2.80] |
| Adjacent, shared dense | 11.07 | [8.67, 13.34] |
| Scaled adjacent, shared dense | 10.54 | [8.27, 12.67] |
| Residual, shared dense | 7.20 | [4.93, 9.20] |
| Residual, coordinate thresholds | -1.33 | [-2.13, -0.67] |
| Residual, greedy stage thresholds | -1.33 | [-2.13, -0.67] |
| Ridge, two checks | 0.40 | [-0.80, 1.60] |
| Gaussian MC, shared dense | 9.94 | [7.67, 12.01] |
| Gaussian MC, two checks | -1.80 | [-3.80, -0.16] |

The patient-paired bootstrap uses random seed 20260921. Each bootstrap sample resamples patients and retains all paired policy outcomes for each selected patient. All discordance counts for the primary failure endpoint are zero because every frozen policy has zero such failures. The saving intervals have marginal rather than simultaneous coverage. The main comparison against fixed $k = 10$ was specified before test opening. All policy contrasts in the table are reported, including those with stronger baselines. These are separate descriptive comparisons, without simultaneous coverage or a risk-noninferiority claim; zero/zero failure counts do not establish equality of population risks.

## F ADC AND SECONDARY QUALITY

Secondary quality analyses keep the frozen stopping decisions unchanged. They vary the discrepancy criterion or summarize ADC errors relative to the full-budget reference.

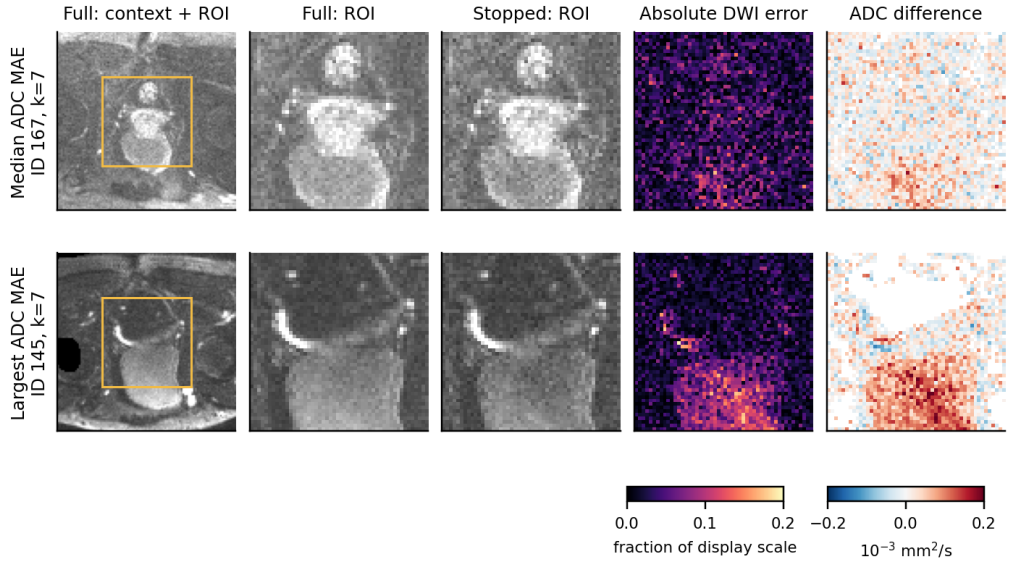



Figure 4: Examples selected by median and maximum patient ADC MAE, showing each patient's slice with largest MAE. Yellow square: 100 mm proxy region. Intensity and error scales are shared across rows; ADC differences use $\pm 0.2 \times 10^{-3}$ mm$^2$/s. Selection and display details are given below.

Figure 4 displays the geometric trace across directions; the consistency endpoint is evaluated separately in each direction. Full-budget and stopped images share the displayed full-budget context's 99th-percentile intensity scale, with the same relative DWI error scale in both rows. The shared ADC color range clips 0.23% of displayed masked pixels in the second row. The examples describe reconstruction differences, without assigning lesion status.

Let $Y_d^{\text{low}}$ be the average of all four $b = 50$ images and $Y_{d,k}^{\text{high}}$ the high-$b$ prefix average in the fixed box. The ADC implementation is

$$\text{ADC}_k = \max\left\{0, \frac{1}{950}\frac{1}{3}\sum_d \left[\log(Y_d^{\text{low}} + a) - \log(Y_{d,k}^{\text{high}} + a)\right]\right\}.$$

The common offset is $a = 10^{-5}$ when the mean of the low-$b$ image and the $k = 4$ high-$b$ prefix (averaged equally) is below $10^{-3}$, and $a = 1$ otherwise, adapting the public reconstruction's intensity-scale convention to a prefix-fixed offset. It is fixed at $k = 4$ and shared by stopped and full-budget ADC. The mask is the geometric low-$b$ trace exceeding 5% of its 98th percentile within the box, also fixed before stopping. Neither the offset nor the mask uses the high-$b$ full reference. No per-$b$ normalization is applied. This two-$b$ engineering calculation is not a lesion-specific clinical ADC measurement or a Bland–Altman noninferiority assessment.

Table 5: Secondary quality at frozen stopping decisions ($n = 45$). Failure counts use the primary region at $m = .10$ or whole slices at $m = .15$. ADC columns give across-patient means of bias, MAE, and within-patient 95th-percentile absolute error, in $10^{-6}$ mm$^2$/s.

| Policy | $m = .10$ | Whole | ADC bias | ADC MAE | ADC p95 |
|---|---|---|---|---|---|
| Fixed $k = 8$ (5% training budget) | 42 | 0 | 1.81 | 25.72 | 65.24 |
| Adjacent, shared dense | 0 | 0 | 0.71 | 17.10 | 43.15 |
| Scaled adjacent, shared dense | 0 | 0 | 0.77 | 17.58 | 44.37 |
| Residual, shared dense | 0 | 0 | 1.41 | 20.15 | 51.05 |
| Residual, two checks (primary) | 31 | 1 | 2.47 | 26.20 | 66.50 |
| Residual, coordinate thresholds | 31 | 1 | 2.33 | 27.38 | 69.52 |
| Residual, greedy stage thresholds | 31 | 1 | 2.33 | 27.38 | 69.52 |
| Ridge, two checks | 30 | 1 | 2.40 | 25.85 | 65.58 |
| Gaussian MC, shared dense | 0 | 0 | 1.40 | 20.13 | 51.01 |
| Gaussian MC, two checks | 37 | 1 | 2.73 | 28.12 | 71.44 |

At $m = .20$, all policies have zero secondary failures. Fixed $k = 10$ has zero failures under both tabulated secondary criteria and ADC MAE 15.94; the full-budget reference has zero discrepancy by construction. The within-patient 95th percentile of absolute error summarizes image discrepancy, not statistical uncertainty.

The examples in Figure 4 are selected after testing by a deterministic descriptive rule: sort patients by primary-policy ADC MAE (ties by ID), display the median-ranked and maximum-ranked patient, and within each choose the slice with the largest masked ADC MAE. These are IDs 167 and 145, both stopped at stage 7. Patient MAEs are 28.91 and 37.29 in $10^{-6}$ mm$^2$/s. This selection illustrates the middle and upper end of patient ADC errors. The mean of patient MAEs need not equal their median.

## G TIMING AND DELIVERY SENSITIVITY

Table 6: Delivery sensitivity of frozen policies (saving, %). Columns $\ell$ deduct image-domain computation plus $\ell$ seconds per visited check. Extra repeat reports acquisition-only saving after one additional high-$b$ repetition per direction following an adaptive stop request, capped at 12. Columns are separate hypothetical scenarios.

| Policy | $\ell = 0$ | $\ell = .25$ | $\ell = 1$ | $\ell = 2$ | Extra repeat |
|---|---|---|---|---|---|
| Fixed $k = 10$ (zero training failures) | 12.00 | 12.00 | 12.00 | 12.00 | 12.00 |
| Fixed $k = 8$ (5% training budget) | 24.00 | 24.00 | 24.00 | 24.00 | 24.00 |
| Residual, shared dense | 17.19 | 16.67 | 15.12 | 13.04 | 11.20 |
| Residual, two checks (primary) | 24.39 | 24.28 | 23.95 | 23.51 | 18.40 |
| Residual, greedy stage thresholds | 25.73 | 25.53 | 24.92 | 24.11 | 19.73 |
| Ridge, two checks | 23.99 | 23.88 | 23.54 | 23.09 | 18.00 |
| Gaussian MC, shared dense | 14.45 | 13.93 | 12.38 | 10.30 | 11.20 |
| Gaussian MC, two checks | 26.20 | 26.10 | 25.80 | 25.40 | 20.80 |

As in Section 5, the main model pauses acquisition for online image-domain processing. For each policy, a decision is counted only at a checkpoint with a finite threshold that is actually visited before stopping. Feature policies are charged for every streaming update through their final visited checkpoint, plus feature computation at the visited checkpoints and actual threshold or ridge calculations. Gaussian MC is charged for its visited draws, sufficient statistics, and comparisons. Offline computation of reference losses is excluded for all policies. CPU measurements use an Intel i9-10900X with two BLAS threads after ingestion has finished, without concurrent ingestion/reconstruction tasks. Each patient's execution is timed once; the reported intervals do not quantify repeated hardware-benchmark variability. Costs after committing to full acquisition are not charged. The hypothetical per-check delays are $\ell \in \{0, 0.25, 1, 2\}$ seconds.

Fixed budgets have no adaptive overshoot. Every policy has zero primary failures on the delivered images in the extra-repeat scenario. These are alternative delivery models for the frozen policies. A controller trained for latency could request a stop earlier, requiring its own development and validation. Scanner-side measurements are needed to select and validate an appropriate delivery model.

## H RISK BOUNDS AND DATA REUSE

For a fixed policy evaluated on $n$ independent patients with common failure probability $p$, the one-sided Clopper–Pearson upper bound is $\mathrm{Beta}^{-1}(0.95; x + 1, n - x)$ for $x < n$ failures, and 1 for $x = n$. At $x = 0$, it is $1 - 0.05^{1/n}$. With $n = 45$, this is 6.4404%. At least 59 zero-failure patients are needed for this single-policy upper bound to be at most 5%. Inspecting several policies and selecting one afterward does not automatically inherit an unadjusted selection guarantee. Risk-control procedures such as Learn then Test (Angelopoulos et al., 2025) address selection under their respective assumptions. No fresh independent risk-calibration stage is used here.

Historical data informed the region, event, and score choices, so they remain development data throughout. The original 45-patient test evaluates policies fixed before its outcomes were opened. Later analyses retain their exploratory status even when their code and inputs are separately recorded. The contiguous ID blocks, incomplete archive tails, single-source protocol, and absence of clinical annotations further limit population interpretation. The empirical failure endpoint concerns consistency with the reconstructed full-budget reference on the retained complete patients.

## I POST-TEST EXPLORATORY SCORE CONTROLS

These analyses were designed after the original 12-policy test was opened. They use the same 107 development patients and separately generated synthetic trajectories to examine covariance estimation and temporal model assumptions. The 45 test patients had also been accessed in subsequent explorations, so they are unavailable as independent confirmation for new scores. This appendix

reports development comparisons only; it neither replaces the frozen primary policy nor changes Table 1.

**Covariance scores.** Suppress direction and slice indices and let $\widehat{\Sigma}_k$ be the unbiased prefix covariance, $\nu = k - 1$, $t_1 = \operatorname{tr}\widehat{\Sigma}_k$, and $t_2 = \operatorname{tr}(\widehat{\Sigma}_k^2)$. For $k < R$, the compared radius scores use

$$\mathcal{Q}_f\left(\frac{\sqrt{\gamma_k t_1 F^{-1}_{\widehat{\eta},\nu\widehat{\eta}}(0.9)}}{\|\bar{X}_k\|_2 + \epsilon}\right), \qquad \gamma_k = \frac{R-k}{Rk}.$$

Here $F^{-1}_{a,b}$ is an $F$-distribution quantile and $\widehat{\eta}$ is an effective-rank estimate, computed separately for each direction and slice. The sample effective rank is $t_1^2/t_2$; it need not equal matrix rank. A moment correction uses

$$\widehat{v}_2 = \frac{\nu^2 t_2 - \nu t_1^2}{(\nu-1)(\nu+2)}, \qquad \widehat{\eta} = \frac{t_1^2}{\operatorname{clip}(\widehat{v}_2, t_1^2/P, t_1^2)}.$$

Under IID Gaussian repeats, $\widehat{v}_2$ is unbiased for $\operatorname{tr}(\Sigma^2)$: the Wishart identities are $\mathbb{E}t_2 = (1 + 1/\nu)\operatorname{tr}(\Sigma^2) + (\operatorname{tr}\Sigma)^2/\nu$ and $\mathbb{E}t_1^2 = (\operatorname{tr}\Sigma)^2 + 2\operatorname{tr}(\Sigma^2)/\nu$. Clipping and taking a ratio do not preserve unbiasedness. Inserting estimated rank into an $F$ quantile is a score approximation, not a coverage theorem.

For OAS and Rao–Blackwell Ledoit–Wolf (RBLW) shrinkage (Chen et al., 2010), use the covariance $(1-\lambda)\widehat{\Sigma}_k + \lambda t_1 I/P$ with

$$\lambda_{\mathrm{OAS}} = \min\left\{1, \frac{(1-2/P)t_2 + t_1^2}{(\nu+1-2/P)(t_2 - t_1^2/P)}\right\},$$
$$\lambda_{\mathrm{RBLW}} = \min\left\{1, \frac{(\nu-2)t_2/\nu + t_1^2}{(\nu+2)(t_2 - t_1^2/P)}\right\}.$$

The effective rank of this shrunk covariance enters the same radius score. The sample size in these formulas is $\nu$: Gaussian Helmert contrasts give $k-1$ independent zero-mean vectors whose covariance estimate is $\widehat{\Sigma}_k$. We retain the original $2/P$ terms. Zero spread gives full shrinkage; zero variance gives zero radius. These baselines compare covariance estimates within one score construction.

The linear-drift control fits $X_r = \mu + r\theta + \xi_r$ by prefix least squares, where $\theta \in \mathbb{R}^P$ is a drift vector and $\xi_r$ is noise. With $h_k = k(k^2-1)/12$, $a_k = (R-k)/2$, and $\widehat{\sigma}^2_{\mathrm{tr}} = ((k-1)t_1 - h_k\|\widehat{\theta}\|^2)/(k-2)$, the estimated energy of the prefix-to-reference difference is $a_k^2\|\widehat{\theta}\|^2 + (\gamma_k - a_k^2/h_k)\widehat{\sigma}^2_{\mathrm{tr}}$. The implementation clips estimated noise and energy below at zero, takes the square root, normalizes by the prefix norm, and applies $\mathcal{Q}_f$.

**Development comparisons.** All methods use the original endpoint and zero training failures. We compare two checkpoints with a shared threshold and the greedy stage-specific thresholds. This follow-up enumerates all distinct training scores at the enabled checkpoints, plus the never-stop option, rather than the original 129-quantile grid. Three fixed random five-fold partitions (seeds 20260921, 20260922, 20260923) and one five-fold contiguous-ID partition are used, with all fitting confined to the training folds. Repeated partitions reuse 107 patients; they do not create 321 independent observations.

The corrected rank's two-checkpoint advantage over OAS and RBLW is only 0.30 and 0.13 percentage points, respectively. These two-checkpoint controls all fail on patient 013 in each of the four partitions. A prespecified resource rule required at least a one-point mean advantage over both shrinkage controls, no additional failures in any partition, and nonnegative gain in ID-blocked validation. The effect-size requirement failed, so no further 45-patient replay or timing was launched for this follow-up. This is a decision to limit further exploration, not a test of equivalence.

**Model stress tests.** For each model and $P \in \{32, 128, 512\}$, generate 2,000 length-12 trajectories and evaluate prefixes $k \in \{4, 7, 10\}$. Independent Gaussian noise has unit coordinate variance and mean 2. Spatial and temporal AR(1) models use coefficient 0.6 with stationary unit variance. The spiked Gaussian covariance has trace $P$, half in one eigenvalue and the rest equally distributed.

Table 7: Exploratory covariance-score controls ($n = 107$). Savings exclude computation and average three five-fold partitions; failures give counts or ranges across partitions. ID blocks uses contiguous patient-ID folds for two-checkpoint policies; all have one failure.

| Score | Two checks (%) | Failures | ID blocks (%) | Stage-specific (%) | Failures |
|---|---|---|---|---|---|
| Residual | 25.98 | 1 | 25.68 | 26.28 | 1–2 |
| Sample rank | 25.55 | 1 | 25.85 | 26.50 | 1–3 |
| Corrected rank | 28.02 | 1 | 27.93 | 28.54 | 2 |
| OAS | 27.72 | 1 | 27.42 | 28.09 | 1–3 |
| RBLW | 27.89 | 1 | 27.76 | 28.21 | 1–3 |
| Linear drift | 26.95 | 1 | 26.80 | 29.03 | 2–3 |
| Gaussian MC | 27.46 | 1 | 27.25 | 27.76 | 2–3 |

Student noise is $t_5\sqrt{3/5}$. Linear drift adds $0.07r$ at repeat $r$; the late-jump model adds 0.5 from repeat 9 onward. The seed is 20260922. The reported event is that $\|\bar{X}_k - \bar{X}_R\|^2$ exceeds the estimated radius squared, before normalization or patient-level aggregation.

Table 8: Radius exceedance (%) for $P = 512$, $k = 7$, and 2,000 simulated trajectories per model. A nominal 90% radius targets 10% exceedance.

| Model | Sample rank | Corrected rank | OAS | RBLW |
|---|---|---|---|---|
| Isotropic Gaussian | 0.00 | 9.30 | 10.30 | 1.10 |
| Spatial AR(1) | 0.00 | 11.75 | 19.65 | 4.70 |
| Spiked covariance | 7.30 | 12.95 | 18.80 | 15.60 |
| Student $t_5$ | 0.00 | 8.65 | 12.50 | 1.40 |
| Temporal AR(1) | 100.00 | 100.00 | 100.00 | 100.00 |
| Linear drift | 0.00 | 100.00 | 100.00 | 100.00 |
| Late jump | 0.00 | 99.95 | 100.00 | 99.95 |

The correction improves scale estimation in several independent-repeat models, but does not address temporal dependence. For the specified temporal AR(1) model at $k = 7$, the exact per-coordinate variance of $\bar{X}_k - \bar{X}_R$ is 0.130884, whereas the expected IID prediction $\gamma_k t_1/P$ is 0.040094. The resulting 3.264-fold mismatch arises before effective-rank estimation. Thus spatial covariance correction cannot by itself justify a stopping guarantee under a misspecified temporal model.

Five development cases (001, 013, 037, 081, 129) were fixed for descriptive diagnostics. At $k = 7$, the largest slice/direction ratio of squared prefix-to-reference error to its IID-predicted expectation is 4.16 for patient 013 and 12.94 for patient 129, despite median ratios near 1.06. This identifies a discrepancy between typical and tail behavior; it does not diagnose motion or determine which slices cause the patient-level event. All patients remain in the analyses.

## J Post-test mechanism experiments

These analyses were specified and executed after the original 45-patient results were inspected. They use only the historically accessed 107 prostate patients and, separately, 25 historically accessed M4Raw brain subjects. The original 12 policies, their code, and their 45-patient outcomes are unchanged. No new policy is fitted or evaluated on those 45 patients here. All new savings are acquisition-only quantities; no new processing times or scanner savings are inferred.

**Prostate design and computation.** We reuse the verified discrepancy trajectories and the residual, Gaussian MC, and OAS score trajectories at stages 4–12. OAS uses the original shrinkage coefficient and the common approximate $F_{.90}$ mapping in App. I; its role is a standard covariance control (Chen et al., 2010). For each score and $\beta \in \{-2, -1, 0, 1, 2\}$, multiply stage $k$ by $\exp(\beta(k - 7.5)/3.5)$. These factors are fixed independently of labels and are not selected for a favorable outcome. Shared-threshold fitting enumerates all distinct training values at the enabled stages, plus the never-stop option; two-checkpoint fitting enumerates all 28 pairs. Maximal integer total saved repetitions determines the optimum, with lexicographic checkpoint and smallest-threshold tie breaking. Stage-specific thresholds use Proposition 1. The $\beta = 0$ two-checkpoint and stage-specific decisions reproduce the earlier exact-grid development comparison for all three scores.

We retain three random five-fold partitions (seeds 20260921, 20260922, 20260923) and a fourth partition consisting of contiguous subject-ID blocks. Fitting uses only the training portion of each partition. We verify every stage's ranking and ties after scaling and every held-out stage-specific decision against its unscaled counterpart. The latter gives 6,420 patient/score/scale/partition checks, all identical. These repeated partitions assess sensitivity to fitting splits; the sampling units remain the same 107 patients.

Table 9: Prostate scale intervention ($n = 107$), fitted with zero training failures. Entries: acquisition saving (%) and out-of-fold failure count. Random columns give mean saving and failure-count range over three five-fold partitions; ID columns use contiguous patient-ID folds.

| Score | $\beta$ | Dense | Two checks | Stage-specific | Dense (ID) | Two (ID) | Stage (ID) |
|---|---|---|---|---|---|---|---|
| Residual | -2 | 17.12 (1) | 25.57 (1) | 26.28 (1–2) | 17.05 (1) | 25.46 (1) | 26.13 (2) |
| Residual | -1 | 17.12 (1) | 25.57 (1) | 26.28 (1–2) | 17.05 (1) | 25.46 (1) | 26.13 (2) |
| Residual | 0 | 18.22 (1) | 25.98 (1) | 26.28 (1–2) | 18.22 (1) | 25.68 (1) | 26.13 (2) |
| Residual | 1 | 0.00 (0) | 22.62 (1) | 26.28 (1–2) | 0.00 (0) | 22.43 (1) | 26.13 (2) |
| Residual | 2 | 0.00 (0) | 22.62 (1) | 26.28 (1–2) | 0.00 (0) | 22.43 (1) | 26.13 (2) |
| Gaussian MC | -2 | 18.26 (1) | 27.70 (1) | 27.76 (2–3) | 18.28 (1) | 27.36 (1) | 27.64 (2) |
| Gaussian MC | -1 | 18.73 (1) | 27.48 (1) | 27.76 (2–3) | 18.90 (1) | 27.36 (1) | 27.64 (2) |
| Gaussian MC | 0 | 19.29 (1) | 27.46 (1) | 27.76 (2–3) | 19.40 (1) | 27.25 (1) | 27.64 (2) |
| Gaussian MC | 1 | 19.31 (1) | 27.48 (1) | 27.76 (2–3) | 19.57 (2) | 27.36 (1) | 27.64 (2) |
| Gaussian MC | 2 | 18.11 (1) | 27.14 (1) | 27.76 (2–3) | 18.17 (1) | 27.25 (1) | 27.64 (2) |
| OAS | -2 | 17.53 (1) | 27.64 (1) | 28.09 (1–3) | 17.50 (1) | 27.31 (1) | 27.81 (2) |
| OAS | -1 | 18.56 (1) | 28.09 (1) | 28.09 (1–3) | 18.62 (1) | 27.53 (1) | 27.81 (2) |
| OAS | 0 | 19.53 (1) | 27.72 (1) | 28.09 (1–3) | 19.68 (1) | 27.42 (1) | 27.81 (2) |
| OAS | 1 | 0.00 (0) | 26.07 (1) | 28.09 (1–3) | 0.00 (0) | 25.51 (1) | 27.81 (2) |
| OAS | 2 | 0.00 (0) | 26.07 (1) | 28.09 (1–3) | 0.00 (0) | 25.51 (1) | 27.81 (2) |

**Interpreting score reversal.** At $\beta = 0$, OAS has higher mean two-checkpoint saving than Gaussian MC; at $\beta = 1$, Gaussian MC has higher saving. This ordering also reverses in the ID-blocked partition: 27.42 versus 27.25% at $\beta = 0$, and 25.51 versus 27.36% at $\beta = 1$. Each of these policies has one out-of-fold failure in the corresponding partition. The transformation changes neither score's within-stage ranking, although it does change cross-stage numerical comparability. A transformed Gaussian MC value is a decision score, not a probability. The reversal is a descriptive comparison of means: it demonstrates sensitivity to cross-stage scale without establishing equal population risks or probability calibration.

**Matched-budget comparisons.** We fit fixed policies and both dense and two-checkpoint shared-threshold policies at $\alpha_{\mathrm{dev}} \in \{0, .025, .05, .10\}$. With $n_{\mathrm{dev}}$ training patients in a fold, the allowable failure count is $\lfloor n_{\mathrm{dev}} \alpha_{\mathrm{dev}} \rfloor$. The fixed rule chooses the earliest stage satisfying that same constraint. The score, endpoint, candidate stages, partitions, and acquisition denominator are unchanged. The stage-specific greedy thresholds are omitted from the positive budget curves because their exact optimality result requires zero training failures. Table 10 gives every fitted comparison, including cases with more out-of-fold failures. No risk-equivalence test is implied.

The hindsight calculation uses $T_i^\star$ from Section 6.2: each patient's first passing candidate stage, or the full-budget fallback, identified from the completed trajectory without a stopping score. There are no pass-to-fail transitions across later stages in these 107 observed primary-endpoint trajectories, though neither the fitting algorithm nor the theoretical arguments require this property. Its 33.14% saving is a ceiling only for policies with zero failures on these realized trajectories. Policies permitting errors may exceed it, and unseen-patient prefix-only policies need not approach it. Policies with out-of-fold failures and this zero-failure hindsight ceiling answer different constrained problems.

Table 10: Prostate comparisons under matched training failure budgets ($n = 107$). Entries: acquisition saving (%) and out-of-fold failure count. Random means, count ranges, and ID folds follow Table 9. The budget constrains training failures.

| Score | Policy | Training budget (%) | Random folds | ID blocks |
|---|---|---|---|---|
| Fixed | Fixed | 0 | 12.00 (0) | 12.00 (0) |
| Fixed | Fixed | 2.5 | 24.00 (2) | 24.00 (2) |
| Fixed | Fixed | 5 | 25.59 (4–7) | 25.18 (4) |
| Fixed | Fixed | 10 | 30.00 (7) | 30.00 (7) |
| Residual | Dense | 0 | 18.22 (1) | 18.22 (1) |
| Residual | Dense | 2.5 | 29.51 (3) | 29.44 (3) |
| Residual | Dense | 5 | 29.87 (5–6) | 29.89 (6) |
| Residual | Dense | 10 | 31.12 (10–12) | 31.07 (11) |
| Residual | Two checks | 0 | 25.98 (1) | 25.68 (1) |
| Residual | Two checks | 2.5 | 29.38 (3) | 29.38 (3) |
| Residual | Two checks | 5 | 29.55 (5–7) | 29.55 (4) |
| Residual | Two checks | 10 | 30.11 (8–10) | 30.34 (9) |
| Gaussian MC | Dense | 0 | 19.29 (1) | 19.40 (1) |
| Gaussian MC | Dense | 2.5 | 29.35 (3–4) | 29.44 (4) |
| Gaussian MC | Dense | 5 | 29.93 (5–6) | 29.94 (4) |
| Gaussian MC | Dense | 10 | 31.08 (9–11) | 31.23 (11) |
| Gaussian MC | Two checks | 0 | 27.46 (1) | 27.25 (1) |
| Gaussian MC | Two checks | 2.5 | 29.10 (3) | 28.93 (3) |
| Gaussian MC | Two checks | 5 | 29.68 (5–7) | 29.38 (4) |
| Gaussian MC | Two checks | 10 | 30.21 (10–11) | 30.28 (12) |
| OAS | Dense | 0 | 19.53 (1) | 19.68 (1) |
| OAS | Dense | 2.5 | 29.33 (2–4) | 29.38 (3) |
| OAS | Dense | 5 | 30.34 (5–7) | 30.22 (5) |
| OAS | Dense | 10 | 30.99 (10–12) | 30.95 (10) |
| OAS | Two checks | 0 | 27.72 (1) | 27.42 (1) |
| OAS | Two checks | 2.5 | 29.08 (3) | 28.93 (3) |
| OAS | Two checks | 5 | 29.35 (3–6) | 29.89 (6) |
| OAS | Two checks | 10 | 30.34 (10–11) | 29.66 (7) |

### J.1 Exploratory M4Raw brain replication

The M4Raw v1.6 release supplies 25 subjects with six repetitions for both T1 and T2 (Lyu et al., 2023). All are included by repetition-count eligibility, without selecting on score or quality. These subjects have been used during earlier project development and are not a new held-out cohort. We read the released $18 \times 256 \times 256$ root-sum-of-squares (RSS) volumes, average magnitude images without a trained denoiser, and use the six-repeat mean as the reference. The whole slice is the region. T1 and T2 are analyzed separately using the same subject partitions; they do not constitute 50 independent subjects.

At each $k \in \{2, 3, 4, 5\}$, the residual score is Eq. 6 with $R = 6$, one contrast instead of three directions, and the second-largest slice value because $\lfloor .1 \times 18 \rfloor + 1 = 2$. A subject-level failure occurs if more than 10% of slices exceed the chosen relative discrepancy tolerance. The brain protocol fixed 0.15 as its primary tolerance and 0.05 and 0.10 as sensitivity analyses before this exploratory run. Saving is $100(6 - k)/6$ percent in equal-repeat units. It is not directly comparable to the prostate 50-entry percentages and is not a measured scanner duration.

The intervention is $z^{(\beta)}(k) = \exp(\beta(k - 3.5)/1.5)z(k)$ at the same five values of $\beta$. We enumerate all six two-checkpoint pairs or all four dense checks, with exact empirical threshold search and zero training failures. The three seeded five-fold partitions and one contiguous-ID partition each contain all 25 subjects. All 3,000 out-of-fold stage-specific subject/contrast/tolerance/scale/partition comparisons are unchanged.

At $m = .15$, T1 dense shared-threshold rules have 1–2 out-of-fold failures at $\beta = 0$ and one at $\beta = 1$ across random partitions; T2 has 0–2 at both scales. T1 stage-specific rules have two failures in every random partition; T2 has 0–2. Thus preserved decisions do not imply zero failures on out-of-fold subjects or establish a generalization advantage for stage-specific thresholds. The ID-blocked partition reproduces the large dense shared-threshold scale effect (T1: 48.00 to 10.67%; T2: 54.00 to 16.00%),

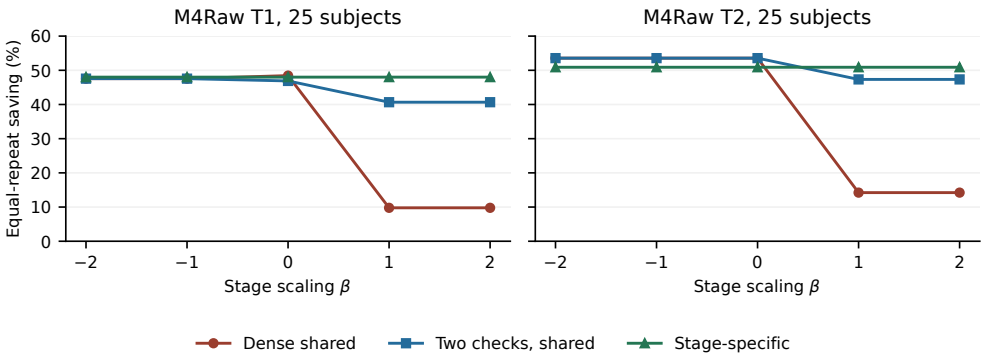



Figure 5: Score-scale sensitivity in brain MRI ($n = 25$, $m = .15$). Savings use six equal-cost repetitions and average three five-fold partitions. T1 and T2 are analyzed separately. Failure counts and other tolerances: Tables 11–12.

with one failure for both scales in each contrast. The other tolerances are included in full to show the scope of the effect, including degenerate settings in which transformed shared policies fall back to the full budget.

Table 11: M4Raw T1 scale intervention ($n = 25$). Entries: equal-repeat saving (%) and out-of-fold failure count. Random means, count ranges, and ID folds follow Table 9. Primary tolerance: $m = .15$.

| $m$ | $\beta$ | Dense | Two checks | Stage-specific | Dense (ID) | Two (ID) | Stage (ID) |
|---|---|---|---|---|---|---|---|
| 0.05 | -2 | 10.89 (1) | 10.89 (1) | 10.89 (1) | 11.33 (1) | 11.33 (1) | 11.33 (1) |
| 0.05 | -1 | 10.89 (1) | 10.89 (1) | 10.89 (1) | 11.33 (1) | 11.33 (1) | 11.33 (1) |
| 0.05 | 0 | 10.89 (1) | 10.89 (1) | 10.89 (1) | 11.33 (1) | 11.33 (1) | 11.33 (1) |
| 0.05 | 1 | 0.00 (0) | 0.00 (0) | 10.89 (1) | 0.00 (0) | 0.00 (0) | 11.33 (1) |
| 0.05 | 2 | 0.00 (0) | 0.00 (0) | 10.89 (1) | 0.00 (0) | 0.00 (0) | 11.33 (1) |
| 0.10 | -2 | 35.78 (0–1) | 35.78 (0–1) | 34.89 (1–2) | 36.00 (1) | 36.00 (2) | 35.33 (2) |
| 0.10 | -1 | 35.78 (0–1) | 35.78 (0–1) | 34.89 (1–2) | 36.00 (1) | 36.00 (2) | 35.33 (2) |
| 0.10 | 0 | 32.44 (1) | 33.11 (0–2) | 34.89 (1–2) | 32.67 (1) | 34.67 (1) | 35.33 (2) |
| 0.10 | 1 | 1.78 (0–1) | 32.00 (0) | 34.89 (1–2) | 2.67 (1) | 32.00 (0) | 35.33 (2) |
| 0.10 | 2 | 1.78 (0–1) | 32.00 (0) | 34.89 (1–2) | 2.67 (1) | 32.00 (0) | 35.33 (2) |
| 0.15 | -2 | 47.56 (1–2) | 47.56 (1–2) | 48.00 (2) | 47.33 (1) | 47.33 (1) | 48.67 (2) |
| 0.15 | -1 | 47.56 (1–2) | 47.56 (1–2) | 48.00 (2) | 47.33 (1) | 47.33 (1) | 48.67 (2) |
| 0.15 | 0 | 48.44 (1–2) | 46.89 (1) | 48.00 (2) | 48.00 (1) | 47.33 (1) | 48.67 (2) |
| 0.15 | 1 | 9.78 (1) | 40.67 (1) | 48.00 (2) | 10.67 (1) | 42.00 (1) | 48.67 (2) |
| 0.15 | 2 | 9.78 (1) | 40.67 (1) | 48.00 (2) | 10.67 (1) | 42.00 (1) | 48.67 (2) |

Table 12: M4Raw T2 scale intervention for the same 25 subjects. Units and conventions follow Table 11. T1 and T2 results are analyzed separately.

| $m$ | $\beta$ | Dense | Two checks | Stage-specific | Dense (ID) | Two (ID) | Stage (ID) |
|---|---|---|---|---|---|---|---|
| 0.05 | -2 | 15.11 (0–2) | 15.11 (0–2) | 15.78 (1–3) | 14.00 (0) | 14.00 (0) | 14.67 (1) |
| 0.05 | -1 | 15.11 (0–2) | 15.11 (0–2) | 15.78 (1–3) | 14.00 (0) | 14.00 (0) | 14.67 (1) |
| 0.05 | 0 | 15.11 (0–2) | 15.11 (0–2) | 15.78 (1–3) | 14.00 (0) | 14.00 (0) | 14.67 (1) |
| 0.05 | 1 | 0.00 (0) | 2.22 (1–3) | 15.78 (1–3) | 0.00 (0) | 1.33 (1) | 14.67 (1) |
| 0.05 | 2 | 0.00 (0) | 2.22 (1–3) | 15.78 (1–3) | 0.00 (0) | 1.33 (1) | 14.67 (1) |
| 0.10 | -2 | 31.78 (1) | 32.67 (1–4) | 35.78 (2–4) | 32.00 (1) | 32.00 (1) | 36.00 (2) |
| 0.10 | -1 | 31.78 (1) | 32.67 (1–4) | 35.78 (2–4) | 32.00 (1) | 32.00 (1) | 36.00 (2) |
| 0.10 | 0 | 32.00 (1) | 32.89 (1–4) | 35.78 (2–4) | 32.00 (1) | 32.00 (1) | 36.00 (2) |
| 0.10 | 1 | 0.00 (0) | 29.33 (1) | 35.78 (2–4) | 0.00 (0) | 29.33 (1) | 36.00 (2) |
| 0.10 | 2 | 0.00 (0) | 29.33 (1) | 35.78 (2–4) | 0.00 (0) | 29.33 (1) | 36.00 (2) |
| 0.15 | -2 | 53.56 (0–2) | 53.56 (0–2) | 50.89 (0–2) | 54.00 (1) | 54.00 (1) | 52.00 (1) |
| 0.15 | -1 | 53.56 (0–2) | 53.56 (0–2) | 50.89 (0–2) | 54.00 (1) | 54.00 (1) | 52.00 (1) |
| 0.15 | 0 | 53.56 (0–2) | 53.56 (0–2) | 50.89 (0–2) | 54.00 (1) | 54.00 (1) | 52.00 (1) |
| 0.15 | 1 | 14.22 (0–2) | 47.33 (0) | 50.89 (0–2) | 16.00 (1) | 48.00 (0) | 52.00 (1) |
| 0.15 | 2 | 14.22 (0–2) | 47.33 (0) | 50.89 (0–2) | 16.00 (1) | 48.00 (0) | 52.00 (1) |