Target-Specific Verification Surfaces for Cross-Stage Quality Assurance: A Medical Image Segmentation Case Study

Michael Rothrock

michael@roth.rocks · https://michael.roth.rocks

Version 1.0 · DOI: 10.5281/zenodo.20331363 · Companion to Trust Topology *(DOI: 10.5281/zenodo.20292194)*

Zenodo empirical companion note · Not peer reviewed · Not for clinical use

Abstract. In a medical AI pipeline, the reliability question is not only "is the model good enough?" but "is the failure you care about visible to the check you deployed?" Trust Topology makes that question precise. It treats each intermediate artifact as a *verification surface* (the representation a stage exposes together with the deterministic checks writable over it) and gives a constructive rule: when a target is invisible at one surface, more zero-false-positive gates over that same artifact cannot verify it; the remedy is to expose a new surface carrying the missing signal. Reliability then composes across stages by conditional escape: each verifier acts on the cases that survived earlier verifiers, so narrow checks with low-overlap rejection sets can still shrink total pipeline escape.

We work this rule end to end on prostate MRI, liver CT, and kidney CT segmentation. Phase 1 restricts verification to a mask-only channel: binary gland and lesion masks plus eleven learning-free anatomical predicates. This channel is strong artifact QA. On calibration-tail failures, every rejection is a true envelope violation and every violation is caught. Across stages, rejection sets are nearly disjoint, with pooled overlap 1.49% (95% CI [0.51%, 4.30%]), so cross-stage portfolios eliminate failures that same-stage budgets miss.

The same channel then hits its limit. Clinically significant prostate cancer is not encoded in mask geometry alone; its ground truth comes primarily from pathology and mpMRI signal. Anatomical gates enrich the flagged cohort but do not verify csPCa: Stage-2 gates catch $\frac{15}{425} = 3.5\%$ of positive cases, Stage-2$\cup$Stage-4 gates catch $\frac{63}{425} = 14.8\%$, and Stage-2 decision-curve net benefit is negative across the threshold grid.

The constructive test is the payoff. Guided by this diagnostic null, Phase 2 adds one literature-derived ADC intensity surface over the joint mask–intensity representation. Held-out csPCa+ coverage rises from 4.5% with Stage-2 gland-mask gates alone to 29.5% for the combined gates+ADC+C1 stack. The joint C1 predicate fires on six held-out test cases, all csPCa-positive, catching $\frac{6}{44}$ positives with $\frac{0}{108}$ false positives; four of those cancers are missed by both marginal surfaces. This is a finite-sample pilot, not a validated detector. Its purpose is to demonstrate the engineering loop: characterize what escapes, add a surface that exposes the missing signal, measure the new residual, and repeat.

1 Introduction

Monday morning. A radiologist opens the day's worklist: eight prostate MRI cases pre-processed by an AI-assisted pipeline. Each case has been through the same chain: T2-weighted acquisition, gland segmentation, lesion detection, PI-RADS scoring. The radiologist's job is not to second-guess the segmentation model, but instead to catch the cases where the pipeline broke down before the report goes out. Case 3: the gland mask covers half the bladder. Case 7: a lesion sits outside the gland boundary. Case 5: the segmented gland volume is 180 cc, three times the population mean. None of those are model failures in a deep sense; they are pipeline failures, caught by eye in seconds.

This paper asks a surface-engineering question: given a chosen artifact channel, which failures can be rejected with deterministic certainty, which targets remain invisible to that channel, and what surface should be added next? The goal is not to replace radiologist review, but to show how verification surfaces compose: characterize the current escape set, add one target-exposing surface, measure the new residual, and repeat.

Trust Topology gives this pipeline-level view a precise vocabulary. A *verification surface* is the artifact exposed at a pipeline stage together with the deterministic checks writable over it, the gates actually deployed, and the target those gates are meant to catch. A surface aligns with a target when failures land in regions of the artifact space that good cases never occupy. When that happens, a zero-false-positive gate is possible. When the artifact does not expose the target, more gates over the same artifact cannot create the missing information.

We deliberately begin with one restricted representation channel: binary segmentation masks plus deterministic anatomical predicates. Across the Stage-2 gland-mask surface and the Stage-4 lesion-list surface, we ask what this mask-only channel can do. It can reject calibration-tail failures with certainty, by construction. It can compose across stages because the two surfaces catch mostly different artifact failures. But what it cannot do is verify clinically significant prostate cancer at useful zero-false-positive or decision-quality levels, because the csPCa signal lives primarily in pathology and mpMRI intensity channels rather than binary mask geometry.

This leads to the next engineering question: can reliability improve not by changing the model, but by changing the verification surfaces, and does the work of individual gates compound when they are orchestrated into a pipeline?

This is the medical companion to the Trust Topology flagship (Rothrock, 2026a). The flagship establishes the framework on a production coding-agent pipeline; this paper tests whether the same surface-target logic transfers to a structurally different setting: medical image segmentation, a non-revision architecture with public datasets and no shared code or models. The result is generalization evidence for the framework, not a standalone clinical claim. The interpretable cross-stage quality-control pipeline studied here was introduced in an earlier preprint (Rothrock, 2026b); this paper recasts it in verification-surface terms and extends it with cross-stage composition analysis and a constructive ADC surface.

We instrument eleven learning-free gates over two medical verification surfaces. Stage-2 gland-mask thresholds are calibrated on 50 PROMISE12 cases and applied unchanged across PI-CAI prostate models ($N = 1500$ for Bosma22b/Guerbet23; $N = 300$ T2W subset for MONAI/TotalSegmentator). Stage-4 lesion-list gates use fixed anatomical priors or the within-cohort envelope stated in §3. For cross-domain transfer, we reuse the same gate-family construction on MSD Task03 liver ($N = 131$) and KiTS23 kidney ($N = 489$), with domain-specific threshold recalibration and anatomically inapplicable predicates disabled.

Three findings organize the paper. First, calibration-tail gates are exact for their construction-aligned target: every rejection is outside the declared anatomical envelope, and every such envelope violation is caught. Second, cross-stage overlap is small ($\omega \approx 1.5\%$ pooled), so what one stage catches the next mostly does not. Third, the mask-only channel produces a csPCa cohort shift but no clinical decision benefit. That diagnostic null then motivates the constructive extension: adding an ADC-derived intensity surface exposes signal absent from the mask-only channel and produces the predicted refinement gain. The radiologist's gland-shape eye can reshuffle the worklist, but it is not a cancer detector, and neither is its automated version. The engineering question is which surface should be added when the current one cannot see the target.

Composition is the operational lever. Stacking verification surfaces reduces misses when they reject disjoint cases (§8.1–§8.3); and where an added surface is aligned to the target, composition can raise both the purity and the coverage of the flagged set, which §8.5 demonstrates on the clinical csPCa target with an added ADC channel. §11 discusses downstream uses for prospective evaluation.

Scope. This is a methodological study structured in two phases. *Phase 1 (main characterization)* studies one deliberately restricted verification channel: AI-produced binary gland and lesion masks with eleven

learning-free deterministic predicates over those masks. *Phase 2 (exploratory constructive extension)* adds one theory-directed ADC-derived intensity surface (§8.4–§8.5) to demonstrate the framework's representation-product composition mechanism (§2) and its above-marginal-union refinement bonus on a clinical target. The paper is not a SOTA csPCa-detection contribution, not a validated csPCa detector or clinical diagnostic method, and not a deployment recommendation; the Phase-2 ADC result is a finite-sample constructive verification-surface pilot to demonstrate the engineering practice, not a validated diagnostic model. Full mpMRI-aware verification is the natural follow-on contribution (see §12, *The natural next contribution*). We compare this chosen channel against split-conformal cross-model disagreement and intra-model TTA softmax uncertainty to show that the framework's regime structure is visible across different verification-surface choices. The medical case is the empirical anchor; the flagship and its Formal Supplement develop the general framework.

How to read this paper. Practitioners deploying medical-imaging QC: §3 (pipeline and gates) and §4 (headline results table) cover the deployment story. Readers verifying the framework empirically: §5 (exact alignment), §6 (diagnostic-null mismatch), §8 (cross-stage complementarity). Reviewers assessing methodological rigor: §3.4 (statistical methods), §12 (limitations).

Reproducibility. All datasets are public challenge releases (PI-CAI, MSD Task03, KiTS23, PROMISE12); no private or internal data are used. Gate code, calibration thresholds, seeds, gate outputs, and a numerical-claim verifier suite are shared in the paper's public repository and archived with this Zenodo record. The appendices hold the audit tables; the Code and data availability subsection in §3 gives the formal statement.

2 The verification surface, applied

For the convenience of readers approaching this paper without first reading the flagship, this section gives a one-page recap of the concept. The empirical results stand on their own. The framework gives them their interpretation.

The pipeline as a chain of artifacts. What flows through the pipeline is not the patient but a sequence of representations of the patient. Each stage produces an artifact, and each artifact is a strictly smaller view than the one before it:

Stage	Artifact	Represents	Loses
1	T2W image plus DWI/ADC	Tissue contrast in voxels	The patient
2	Gland mask	Whether a voxel is inside the prostate	Tissue contrast
3	Lesion probability field	Voxelwise lesion likelihood	Discrete lesion geometry
4	Lesion mask / list	Discrete lesion geometry and descriptors	Raw image signal and gland-shape detail
5	csPCa score / clinical report	Risk number or recommendation	Spatial structure and calibration details

The gates in this paper cover the Stage-2 gland-mask surface and the Stage-4 lesion-list surface; they see only the binary mask artifacts emitted at those stages. They cannot see artifact 1′s intensity channel, and they cannot see the patient. Every guarantee a gate makes is a guarantee about its artifact, not about the layer above or the patient below.

A *verification surface* at stage k against target violation type α is a tuple

$$\Sigma_{k,\alpha} := \left(Y_k : \Omega \to S_k,\ \mathcal{D}_k,\ F_k^{\text{deploy}},\ L_\alpha\right)$$

where Y_k is the representation map (here, a segmentation mask or lesion list). $\mathcal{D}_k$ is the deterministic predicate class (here, anatomical predicates with calibrated thresholds). F_k^{deploy} is the deployed gate family

(deterministic plus any stochastic or heuristic components). The event $L_\alpha \subseteq \Omega$ is the target violation event. Each deployed gate g produces a *rejection event* $R_{k,g} \subseteq \Omega$ (the cases on which g fires). For deterministic artifact predicates this factors as $R_{k,g} = \{Y_k \in A_g\}$ for some region $A_g \subseteq S_k$.[1] The *target-relative deterministic subfamily* collects deployed deterministic gates whose rejection events lie inside the target:

$$F_{k,\alpha}^{\det} := \left\{ g \in F_k^{\text{deploy}} : R_{k,g} = \{Y_k \in A_g\},\ A_g \in \mathcal{D}_k,\ Y_k^{-1}(A_g) \subseteq L_\alpha \right\}.$$

A gate may be in $F_{k,\alpha}^{\det}$ for one target and not for another. The stage elimination event is $E_{k,\alpha} := \bigcup_{g \in F_k^{\text{deploy}}} R_{k,g}$. Three measurement attributes characterize the surface:

- $\eta_{k,\alpha}^*$: *capacity*, supremum, over zero-false-positive rejection regions, of target-positive mass. In the finite empirical/atomic setting used here, the Surface Capacity characterization identifies this with $P(C_{k,\alpha} \mid L_\alpha)$, where $C_{k,\alpha}$ is the *catchable set*: the target-positive cases that land in pure-positive buckets (buckets whose cases are all target-positive). It is the most a zero-false-positive gate could ever catch.[2]
- $\rho_{\alpha(g)}$: *purity*, $P(L_\alpha \mid R_{k,g})$ with the deterministic specialization $R_{k,g} = \{Y_k \in A_g\}$, the target-positivity rate among a gate's rejections. Membership in $F_{k,\alpha}^{\det}$ requires $\rho_\alpha = 1$.
- $\Gamma_{k,\alpha}$: *target coverage*, observed recall of the full deployed family F_k^{deploy}, including stochastic and heuristic gates that are not in $\mathcal{D}_k$.

A fourth, ω, measures complementarity within a stage or between stages and is the Jaccard index of rejection sets.

Concretely: η bounds what the surface can catch at zero false positives, and ρ is the target-positive rate within a rejection set. Γ is the fraction of target-positive cases the deployed gates flagged, and ω is how much different gates overlap.

Two structural results from the flagship/Formal Supplement carry through to the medical setting. First, *Refinement Monotonicity*: when a representation Y' refines Y (every Y'-bucket lies inside some Y-bucket), pure-positive coarser buckets remain pure under refinement. Mixed coarser buckets may split into pure subbuckets. Second, the *Cross-Stage Chain Rule* gives the following identity for a fixed target α:

$$P(U^{(\alpha)}) = P(L_\alpha) \cdot \prod_{k \in K_\alpha} P\left((E_{k,\alpha})^c \ \mid\ L_\alpha,\ \bigcap_{j<k} (E_{j,\alpha})^c \right)$$

In words: a target-positive case escapes the whole pipeline only if every relevant stage misses it. So its escape probability is the defect prevalence $P(L_\alpha)$ times, for each stage k, the chance that stage misses it ($(E_{k,\alpha})^c$ is the per-stage escape event) given that the earlier stages already did. Because every factor is at most one, escape shrinks fastest when stages catch *different* cases, the multiplicative payoff that lets a pipeline of imperfect gates be reliable. The independent-product approximation is the special case where stage escapes are conditionally independent given L_α; the cross-stage complementarity result in Section 8 (near-disjoint per-stage rejection sets, $\omega \approx 1.5\%$) is its empirical content.

A note on what composes how. These two results govern different mechanisms. Refinement Monotonicity applies to the *granularity* of a representation: a finer representation has buckets that lie strictly inside coarser ones, and pure-positive coarser buckets remain pure under refinement. The Cross-Stage Chain Rule applies to the *temporal sequence* of per-stage escape events along a pipeline. A useful special case combines them: two representations over the *same case space*, with different predicate classes or representation maps Y_A and Y_B, form a joint representation $Y_{\text{AB}} = (Y_A, Y_B)$ that refines each marginal, because every joint bucket lies inside one Y_A bucket and one Y_B bucket. The two representations need not live on the same artifact layer: they can be mask-derived and intensity-derived, for instance, as long as both are evaluated on the same case.

[1] All inclusions in this paper are understood up to P-null sets.

[2] For continuous representations, the Formal Supplement gives the general measurable form of the same quantity: the *singular mass*, the fraction of target-positive cases whose representation value is one that no target-negative case ever produces. This paper works entirely in the atomic setting (discrete predicates and calibrated thresholds), so the atomic form above is the one used throughout.

We call this *representation-product composition*; it is governed by Refinement Monotonicity rather than the chain rule. The chain rule can still factorize escape over the joint surface once rejection events are defined, but the ordering in that factorization is analytic, not pipeline-temporal. Section 8.4 reports a minimal empirical instance over the binary-mask representation (Y_A) and an intensity-aware ADC representation (Y_B) at matched rejection rate; Section 8.5 reports a clinically-grounded joint predicate that demonstrates above-marginal-union refinement on a held-out test split with frozen thresholds after a three-candidate screen.

Distinguishing two related properties. A surface can have $\rho = 1$ (formal purity, by construction) for one target while only $\hat{\rho} < 1$ (empirical purity, finite-sample false positives observed) for a related target. The medical pipeline makes this distinction central: calibration-tail membership is formal $\rho = 1$ by definition. Other targets, including external Dice-disagreement and csPCa+, have empirical $\hat{\rho} < 1$. We mark the distinction throughout.

2.1 Target / metric glossary

The pipeline is evaluated against four main target families (a fifth, broad lesion-containment disagreement, is an appendix sensitivity target; Appendix E). Each target maps onto this view's vocabulary differently, so we summarize them here:

Target	Symbol	Surface aligned?	Main metric	Section
Calibration-tail membership	L_{tail}	yes (Stage 2, by construction)	formal $\rho = \eta = 1$	§5
External Dice disagreement	$L_{\text{Dice}<0.95}$	no / weak	$\hat{\rho}$, $\hat{\Gamma}$	§7
Tumor-present (cross-domain)	L_{tumor}	partial	target-label $\hat{\Gamma}$, greedy delta	§10
csPCa-positive (clinical)	L_{csPCa}	no (mismatched)	$\hat{\Gamma}$, AUROC, decision curve	§6

The first row is an exact-alignment positive: calibration-tail membership is formal $\rho = \eta = 1$ by definition. The remaining empirical target rows require finite-sample measurement. We use $\hat{\eta}$ exclusively when a rejection region is shown to be zero-false-positive for the target; otherwise we use $\hat{\Gamma}$ (target coverage, observed recall of the deployed family) and report $\hat{\rho}$ separately.

3 Pipeline, datasets, and gates

With the framework's surfaces named (§2), this section instantiates them on the medical pipeline: the five stages, the public datasets, and the eleven deterministic gates.

3.1 Pipeline architecture

The primary pipeline is a five-stage prostate cancer detection workflow:

1. **Stage 1, Image acquisition.** PI-CAI bpMRI inputs, including T2W volumes and diffusion-derived ADC maps. Phase 1 verification consumes only downstream binary mask artifacts; Phase 2 (§8.4–§8.5) explicitly introduces ADC-derived verification predicates.
2. **Stage 2, Gland segmentation.** Whole-prostate mask from a learned model (Bosma22b (Bosma et al., 2022), Guerbet23 (Debs et al., 2023), MONAI (Cardoso et al., 2022), or TotalSegmentator (Wasserthal et al., 2023)). *Verification surface* S_2.
3. **Stage 3, Lesion candidate generation.** Per-voxel lesion probability field from a learned model (Bosma22a (Bosma et al., 2023)). No verification surface at this stage.
4. **Stage 4, Lesion list / lesion-mask verification.** The final scoring stage consumes a discrete lesion artifact (mask + count + per-lesion descriptors). *Verification surface* S_4.

5. **Stage 5, Patient-level csPCa scoring.** The lesion artifact yields a clinically significant prostate cancer score.

We place verification surfaces at Stages 2 and 4, where the artifact representation (segmentation mask, discrete lesion list) admits deterministic anatomical predicates. We place no surface at Stage 1 (raw MRI) or Stage 3 (per-voxel probability field), because the natural deterministic predicates over those representations do not separate target-positive from target-negative cases at usable purity.[3]

Cross-domain transfer pipelines retain only the two verification-surface stages (no clinical-outcome target):

- **Liver tumor (MSD Task03, $N = 131$).** S_2: liver mask. S_4: liver tumor mask.
- **Kidney tumor (KiTS23, $N = 489$).** S_2: kidney mask. S_4: kidney tumor mask.

3.2 Datasets

Dataset	Modality	N	Target+	Notes
PI-CAI (Saha et al., 2024)	MRI (T2W)	1,500	425 csPCa+	Prostate cancer challenge; clinical outcome from `marksheet.csv`
PROMISE12 (Litjens, 2014)	MRI (T2W)	50	—	Calibration-only set; expert gland segmentations
MSD Task03 (Antonelli et al., 2022)	CT	131	118 tumor+	Liver and tumor segmentation
KiTS23 (Heller et al., 2023)	contrast-enhanced CT (corticomedullary or nephrogenic)	489	489 tumor+	Kidney + tumor + cyst segmentation

PROMISE12 is used exclusively for gate-threshold calibration; no PROMISE12 case appears in any evaluation set. PI-CAI, MSD, and KiTS23 evaluation sets are entirely distinct from PROMISE12.

Expert annotations on every case, with two anchor types. All four datasets are competition or grand-challenge releases with matched annotations on every case. The expert anchors are PROMISE12 expert gland segmentations (calibration only), PI-CAI expert lesion annotations released with the PI-CAI labels package (Saha et al., 2024), and PI-CAI clinical csPCa+ outcomes from the patient marksheet. They also include MSD Task03 expert liver/tumor masks and KiTS23 expert kidney/tumor/cyst masks. We pin the exact label state (DIAGNijmegen, 2025): `DIAGNijmegen/picai_labels` at commit `c6bab0b` (tag `v2.0-37-gc6bab0b`, 2025-07-01, which adds expert-derived lesion annotations for previously AI-only positive cases). Expert lesion analyses use `csPCa_lesion_delineations/human_expert/resampled` (available for 1,295 cases) plus `csPCa_lesion_delineations/human_expert/Pooch25` where applicable. `Pooch25` is the 205 positive-case expert annotations added in the 2025-07-01 update (Pooch et al., 2025), so expert-derived lesion annotations are available for all 1,500 cases; each case records its `expert_label_source`. AI-derived Bosma22a lesion masks (`csPCa_lesion_delineations/AI/Bosma22a`) and Bosma22b / Guerbet23 gland masks (`anatomical_delineations/whole_gland/AI`) are used only where explicitly labeled as model artifacts or strong-reference proxies. Gland-mask comparisons on PI-CAI use these AI-derived strong-reference annotations rather than expert gland labels. PI-CAI does not release per-case expert prostate-gland segmentations on the 1,500-case evaluation cohort. Lesion-containment, calibration-tail, low-Dice cross-model disagreement, ISUP severity stratification, and csPCa-outcome comparisons in this paper use PI-CAI expert lesion/clinical labels or the AI-derived strong-reference proxy. Each analysis labels which anchor it uses.

3.3 The eleven gates

Stage 2 gates (8 anatomical predicates over gland masks):

[3]Throughout the paper we use the labels "Stage 2 / S_2" and "Stage 4 / S_4" to refer to the gland-mask and lesion-list verification surfaces respectively.

1. `volume_bounds`, gland volume falls outside the calibrated $[p_{10}, p_{90}]$ envelope.
2. `single_component`, mask has more than one connected component.
3. `convexity`, convex-hull-to-mask volume ratio outside calibration envelope.
4. `centroid_smoothness`, slice-to-slice centroid displacement outside envelope.
5. `slice_area`, per-slice area falls outside envelope.
6. `slice_continuity`, gap between consecutive nonempty slices.
7. `aspect_ratio`, bounding-box aspect ratio outside envelope.
8. `anatomical_location`, mask location outside expected anatomical region.

Stage 4 gates (3 lesion-anatomical predicates):

9. `lesion_in_gland`, more than 50% of lesion volume outside the gland mask.
10. `lesion_volume`, lesion volume outside the calibrated $[p_{10}, p_{90}]$ envelope.
11. `lesion_count`, number of lesion components outside envelope.

Stage-2 calibration source. We calibrate Stage-2 thresholds as p_{10}/p_{90} percentiles on the 50 PROMISE12 expert gland segmentations. *Within the prostate domain*, we apply these thresholds without modification across all four PI-CAI segmentation models.

Stage-4 calibration source. The deployed `lesion_in_gland` threshold ($> 50\%$ of lesion volume outside the organ) is an anatomical prior, not data-derived: only gross containment failure triggers rejection, since smaller spillover reflects ordinary boundary and registration disagreement between the two mask sources. The `lesion_count` envelope (≤ 10 connected components) is likewise a fixed literature-informed bound. The `lesion_volume` envelope (≤ 50 cc) is different: an initial literature-informed ceiling was loosened against this cohort's expert lesion annotations, so we treat it as within-cohort calibration rather than held-out validation. That gate is nearly inert, rejecting one case in 1,500, and produced no observed false rejection of valid focal lesions in the audit; the Stage-4 rejection signal is carried by the fixed-prior `lesion_in_gland` gate. Appendix A gives the full calibration audit. The empirical-purity table in §7 reports Stage-2 measurements.

Cross-domain. Liver and kidney pipelines reuse the same gate logic and two-stage architecture with domain-specific threshold recalibration. We disable anatomically inapplicable predicates (e.g., gland-shape gates that do not generalize to liver or kidney organs) by domain. Thresholds transfer within domain, not across domain.

3.4 Statistical methods

Confidence intervals on proportions use the Wilson interval (Wilson, 1927). Bootstrap CIs use 10,000 case-resampling iterations, seed = 42. Independence tests for cross-stage co-rejection use Fisher's exact test (Fisher, 1935) with Cochran–Mantel–Haenszel pooling (Cochran, 1954; Mantel & Haenszel, 1959) across domains. Decision-curve analysis (Vickers & Elkin, 2006) evaluates net benefit at clinically relevant threshold probabilities. The AI risk score is calibrated lesion volume (5-fold cross-validated logistic on $\log(1 +$ lesion volume in cc) predicting csPCa+). For decision-curve analysis the calibrated Stage-2 filter is applied as follows. Rejected cases are escalated to human review and excluded from the automated-scoring arm. Net benefit is computed over the original full cohort, with escalated cases scored as not-automatically-flagged. The reported Δ NB is the filtered policy's net benefit minus the unfiltered AI-score policy's net benefit at the same threshold.

3.5 Baselines and scope of comparison

The comparisons in this paper test this approach's central question: does the deployed verification surface separate the target? Six baselines are evaluated against the deterministic gates:

1. *Random matched-rate case rejection* (used in Section 6 for csPCa cohort shift and clinical outcomes). 1,000 random 95-case subsets at the matched rejection rate, seed = 42. Section 8′s greedy-portfolio comparison uses a methodologically distinct random baseline: random gate-portfolio subsets per gate budget b, not matched-rate case rejection.

2. *Threshold-tightening sweep* (Section 6): the same gate logic with stricter p_3/p_{97} envelopes, scaling rejection rate from 6.3% to 16.7%. Tests whether the cohort effect is rate-dependent.
3. *Strong-reference disagreement* (Section 7): Bosma22b vs Guerbet23 Dice on the same case; used as both an external target and as a calibrated rejection classifier.
4. *Single-feature deterministic baselines* (Section 7.2): single-gate variants, `centroid_smoothness` alone, `volume_bounds` alone, `convexity` alone. These hold gate logic constant but reduce the deployed family to one predicate. Tests whether the Stage-2 gate construction adds anything over its strongest individual gate.
5. *Split-conformal on cross-model disagreement* (Section 7.3): formal-coverage baseline using $1 - \mathrm{Dice}(\mathrm{Bosma22b}, \mathrm{Guerbet23})$ as a nonconformity score. 50/50 cal/test split, seed = 42.
6. *Intra-model uncertainty (TTA softmax)* (Section 7.4): single-fold nnU-Net v1 prostate-gland prediction plus three flip augmentations. Per-case scores are mean foreground softmax entropy, TTA Dice variance, and TTA-ensemble entropy. Same 50/50 cal/test split as the conformal baseline. Uses the public BAMF Health nnU-Net v1 prostate-gland release (Zenodo DOI 10.5281/zenodo.8290093), an nnU-Net v1 model on the same prostate-gland T2 MRI task as Bosma22b. We do not claim shared training-data lineage (see §7.4 footnote).

The paper does *not* benchmark against learned QC classifiers (a separately trained quality-control head). The intra-model TTA softmax baseline addresses the standard "model-confidence as QC" reviewer question. Learned QC heads test a different mechanism (an additional supervised model trained for the QC task) and are out of scope here.

3.6 Computational footprint

All eleven gates are pure functions of segmentation masks. They run on CPU in NumPy / SimpleITK with no model loading, no GPU, and no inference cost. Per-case wall-clock cost is dominated by mask I/O. The gate computations themselves are negligible compared to the upstream learned segmentation, which typically requires a GPU and produces the mask in seconds-to-minutes per case depending on model and hardware. For deployment, the gate family adds negligible marginal compute on top of an existing segmentation pipeline.

3.7 Code and data availability

All four source datasets are open public grand-challenge releases (PROMISE12 calibration-only; PI-CAI, MSD Task03, and KiTS23 used for evaluation). They are freely downloadable from the official challenge sites by any researcher after standard registration:

- **PI-CAI** (1,500 prostate MRI cases, expert lesion masks, csPCa labels): pi-cai.grand-challenge.org.
- **PROMISE12** (50 expert prostate gland segmentations, used here as calibration-only): promise12.grand-challenge.org.
- **MSD Task03, Liver** (131 liver CT cases with liver and tumor masks): medicaldecathlon.com.
- **KiTS23** (489 kidney CT cases with kidney, tumor, and cyst masks): kits-challenge.org/kits23.

No private, institutional, or de-identified-internal data is used in this paper. Every numerical result reported here is intended to be reproducible from the public challenge data, public segmentation models or released upstream artifacts, and the gate code, scripts, and numerical-claim verifier suite released in the paper's public repository and archived with this Zenodo record. Gate threshold files, gate implementations, and per-case uncertainty-extraction outputs are released alongside this paper. Random seeds for bootstrap, random-rejection baselines, the conformal calibration split, the TTA inference split, and stratified sampling are stated where they are used (seed = 42 throughout). The Bosma22b / Guerbet23 / MONAI / TotalSegmentator segmentation outputs used as upstream model artifacts are reproducible from the published model weights and the PI-CAI input volumes. The Bosma22b and Guerbet23 outputs are themselves released as part of the PI-CAI challenge artifacts. The intra-model TTA softmax baseline (Section 7.4) uses the BAMF Health public release of the nnU-Net v1 prostate-gland model (Zenodo DOI 10.5281/zenodo.8290093).

4 Results summary

With the pipeline and eleven gates defined (§3), we collect the headline results in one place before the per-target detail:

Target	$\boldsymbol{\rho}$ / $\hat{\boldsymbol{\rho}}$	$\boldsymbol{\eta}$ / $\hat{\boldsymbol{\Gamma}}$	**Headline**	**Reading**
Calibration-tail (S_2)	1 formal	1 formal	exact	surface aligned by construction
Dice < 0.95 vs Guerbet23 (external)	$\hat{\rho} = 6.3\%$	$\hat{\Gamma} = 6.9\%$	low	external strong-reference mismatch
csPCa+ (external clinical), mask-only	$\hat{\rho} = 15.8\%$ among rejects	S_2 $\hat{\Gamma} = 3.5\%$; $S_2 \cup S_4$ $\hat{\Gamma} = 14.8\%$	DCA negative	clinical mismatch (cohort shift only)
csPCa+, exploratory ADC extension (§8.4–§8.5)	C1 finite held-out $\hat{\rho} = 1$ (6/6)	C1 $\hat{\Gamma} = 13.6\%$; gates+ADC+C1 stack $\hat{\Gamma} = 29.5\%$	finite-sample constructive pilot	adding the missing information channel begins to expose the clinical target
Tumor-present (liver, $N = 131$)	$\hat{\rho}$ not central	$\hat{\Gamma} = 3.4\%$	no cross-stage delta	S_2 dormant in this domain/ threshold
Tumor-present (kidney, $N = 489$, all tumor+)	$\hat{\rho}$ not central	$\hat{\Gamma} = 10.2\%$	+0.8 pp at $b = 4, 5$	target-positive coverage, not discrimination

Table 2: Headline results across target families. Calibration-tail is the construction-aligned exact row; external targets are empirical; the ADC row is a finite-sample constructive pilot after a limited three-predicate screen. Cross-stage architectural results are developed in §8.

The rest of the paper develops each row in turn. Section 5 covers the construction-aligned target. Section 6 reports the csPCa diagnostic-null in detail. Section 7 reports empirical purity and external baselines. Section 8 reports cross-stage complementarity and the Phase-2 ADC constructive extension. Section 9 reports model-quality scaling. Section 10 reports cross-domain transfer.

5 Exact alignment: calibration-tail membership

We start at the top of the alignment spectrum, where the framework is sharpest. For one target family the gates make a deterministic guarantee: when they fire, the gland-mask artifact lies inside the declared calibration-tail target (equivalently, outside the calibrated anatomical operating envelope) with certainty. The target is calibration-tail membership, being outside the calibrated anatomical envelope on at least one Stage-2 predicate. Because the target is defined by what the gates check, every gate-rejection is, by definition, a target-positive case: no false positives are possible, and every target-positive case the gate evaluates is caught. (Formally: $\rho = \eta = 1$.)

5.1 Calibration-tail target (S_2)

The Stage-2 calibration-tail target is the most direct illustration of the Surface Capacity characterization. Because tail membership *is* the predicate, the pure-positive set P_{tail} is the union of buckets where at least one gate fires. The catchable set C_{tail} is, by definition, the entire calibration-tail subset of L_{tail}. Therefore both purity and capacity are unity by construction:

$$\rho_{\text{tail}} = 1 \quad \text{(formally, by construction)}$$

$$\eta^*_{\text{tail}} = \hat{\eta}_{\text{tail}} = 1 \quad \text{(formally, by construction)}$$

5.2 Why the exact-alignment positive matters

Tail membership is the one target the gland surface verifies exactly by construction; the tested external targets, Dice disagreement, tumor-present, and csPCa+, are not declared by the gates, so their $\hat{\rho}$ is measured rather than exact.

Stating the positive explicitly does two things. First, it shows what perfect surface-target alignment looks like. Second, it makes the empirical-target results non-trivial. Low $\hat{\rho}$ for an external target is informative precisely because $\rho = 1$ is achievable for an aligned target on the same surface. The contrast forces the analysis to be about target alignment, not gate quality.

Empirical clinical correlation on the calibration-tail subset. Table 2 summarizes how the calibration-tail subset behaves on external clinical outcomes. These are not capacity measurements; they are descriptive statistics of the rejected cohort.

Outcome	**Accepted**	**Rejected**	**Δ (95% CI)**	***p***
Lesion containment fraction	0.785 ($n = 394$)	0.779 ($n = 14$)	+0.006 [−0.142, +0.198]	0.45
csPCa-positive rate	29.2% ($\frac{410}{1405}$)	15.8% ($\frac{15}{95}$)	−13.4 pp [−21.8, −2.4]	0.0072
Lesion FP volume (csPCa−, cc)	0.000 ($n = 995$)	0.000 ($n = 80$)	0.0	1.0

Table 3: Calibration-tail subset clinical descriptive statistics. The 13.4 pp csPCa-rate difference is a cohort effect: tail-rejected cases are systematically lower in clinical positivity. This is a property of the rejected subset, not evidence of detection. Section 6 shows that decision-curve net benefit is *negative* for this filter, formalizing the cohort-vs-detection distinction.

The 13.4 pp csPCa shift is real, but on its own it does not constitute a detection improvement (Section 6). In this view's vocabulary: the calibration-tail surface is target-aligned for tail membership, not for csPCa.

6 Diagnostic-null: csPCa as a mismatched target

Section 5 showed the gland surface at its most distinct: formal $\rho = \eta = 1$ for a target it defines by construction. We now turn the *same* surface on a target it was not built for, whether the patient has csPCa, a *patient-level* property, and ask what happens when artifact-level gates are made to speak to it. The csPCa target tests the framework's diagnostic-null prediction: a clinical-outcome target whose information channel is not in the chosen verification channel.

Here the gates do not make their deterministic-elimination guarantee against csPCa: the cases they remove are not, with certainty, csPCa-positive cases; they are artifact-envelope violations that only weakly correlate with the cancer signal. csPCa+ labels in PI-CAI are primarily histopathology-derived: `case_csPCa` is positive when biopsy histopathology confirms Gleason group ≥ 2 / ISUP ≥ 2, with non-biopsied cases handled by the PI-CAI clinical-label protocol.

Expert lesion delineations come from radiologists reading the full mpMRI study, including T2W and diffusion/ADC channels, often with clinical and pathology context. The deployed gates see only binary segmentation masks, with no access to T2W intensity, diffusion signal, or any imaging information radiologists used. We expect, and measure, the diagnostic null. This is not a failure of the gates; it is the framework's diagnostic value made empirical: the chosen channel tells us in advance that csPCa is out-of-regime for it.

The Surface Capacity characterization provides the formal structure. If the gland-shape representation places few or no csPCa-positive cases in pure-positive buckets, no zero-false-positive deterministic gate over that representation can achieve nonzero recall against csPCa. The deployed gates are evidence about the realized predicate class, not a proof that full measurable capacity is small under all conceivable predicates over the same representation.

Empirically, against L_{csPCa} = csPCa-positive cases ($n = 425$):

Quantity	Estimate	Notes
Stage-2 calibrated csPCa+ coverage $\hat{\Gamma}_{S_2,\text{csPCa}}$	3.5% (15/425)	Stage-2 anatomical-tail union (95 rejections of 1,500). The full Stage-2 $\cup$ Stage-4 deployed-family csPCa coverage is higher, $\frac{63}{425} = 14.8\%$ (derived from the §8 conditional-escape caught counts $c_{S_2} = 15$, $c_{S_4} = 50$, intersection $= 2$ on the 425-case csPCa+ set); this row is the Stage-2 component only.
Calibrated-rejection csPCa+ rate	15.8% vs accepted 29.2% ($\Delta_{\text{csPCa}} =$ rejected $-$ accepted $= -13.4$ pp)	$p = 0.0072$ chi-square Yates (Yates, 1934)
Risk-coverage AUROC, no filter	0.980 ($[0.970, 0.989]$)	Calibrated lesion volume; CV-logistic
Risk-coverage AUROC, calibrated gate	0.980 ($[0.971, 0.989]$)	Identical to no-filter
Decision-curve net benefit, unweighted	$\Delta = -0.0093$ ($[-0.0147, -0.0047]$)	Across threshold grid 5%–50%
Decision-curve net benefit, ISUP-weighted	$\Delta = -0.0240$ ($[-0.0407, -0.0100]$)	ISUP severity weights; ≤ 0 at all thresholds

Table 4: csPCa diagnostic-null measurements. The deployed surface produces a 13.4 pp cohort shift in csPCa+ rate but no detection improvement (residual AUROC unchanged) and negative decision-curve net benefit at all tested thresholds, severity-weighted or not. Both decision-curve CIs lie strictly below zero.

The 13.4 pp csPCa-rate shift on the rejected subset is real (chi-square $p = 0.0072$). But it is a *cohort* effect, not detection. Three independent measurements support the diagnostic-null reading:

1. **Risk-coverage AUROC is unchanged.** Removing the 95 calibrated-gate cases leaves residual AUROC at 0.980, identical to the no-filter baseline. A peak residual AUROC of 0.983 exists, but only at 38% rejection (570 cases), not at the 6.3% calibrated operating point.

2. **Decision-curve net benefit is negative.** Across the threshold grid 5%–50%, Δ NB $= -0.0093$ unweighted with 95% CI $[-0.0147, -0.0047]$. The CI is entirely below zero. ISUP-severity-weighted decision curves are worse: Δ NB $= -0.0240$ with 95% CI $[-0.0407, -0.0100]$.

3. **Negative-control comparison.** Random matched-rate rejection (95 cases drawn uniformly) produces a csPCa-rate difference of 0.2 ± 5.0 pp, much smaller than the 13.4 pp calibrated-gate shift. This supports the reading that the cohort effect is structural. But the calibrated rejection still does not produce detection-axis gain, in contrast to the (random-comparable) negligible gain on lesion containment (0.006 ± 0.051).

Weather-radar gloss. A rainfall radar channel is not a lightning detector merely because its resolution improves. Likewise, mask-geometry gates can be exact for anatomical-tail failures while failing to verify csPCa.

This is the framework's predicted pattern. The deployed Stage-2 surface targets gland-shape anatomical-tail predicates. csPCa positivity is a clinical-outcome target whose dependence on gland shape, at the tested predicate granularity, does not produce useful zero-false-positive coverage in this experiment.

6.1 Severity stratification

For completeness, we stratify the calibrated-gate rejection rate by ISUP severity grade (csPCa+ cases only):

ISUP	n csPCa+	Rejected	Rate	Wilson 95% CI
2	234	8	3.4%	[1.7%, 6.6%]
3	99	2	2.0%	[0.6%, 7.1%]
4	40	3	7.5%	[2.6%, 19.9%]
5	52	2	3.8%	[1.1%, 13.0%]

Table 5: ISUP severity stratification of calibrated rejection rate among csPCa+ cases. ISUP 2 rejection rate (3.4%) vs ISUP 4–5 (5.4%): difference −2.0 pp; Fisher $p = 0.53$. The severity-weighted csPCa+ rejection fraction is 4.1% vs unweighted 3.5%, small enough that the ISUP-weighted decision-curve Δ NB $= -0.0240$ above is dominated by the misalignment, not by clinically harmful preferential rejection of high-severity cases.

The severity-stratified rejection rates do not differ significantly across grades. There is no preferential rejection of low-severity (or high-severity) cases at clinically harmful magnitude. The diagnostic-null reading does not appear to be driven by severity imbalance.

6.2 A dual reading: subgroup identifier vs. decision-quality null

The same numbers admit two distinct clinical readings, both useful in different workflows.

Reading 1, subgroup identifier (positive). The calibrated rejection identifies a subgroup with substantially lower csPCa+ prevalence than baseline:

- Baseline: $\frac{425}{1500} = 28.3\%$ csPCa+.
- Rejected subgroup: $\frac{15}{95} = 15.8\%$ csPCa+ (-44.3% relative risk vs baseline).
- Accepted subgroup: $\frac{410}{1405} = 29.2\%$ csPCa+ ($+3.0\%$ relative enrichment over baseline).
- Odds ratio (csPCa+, rejected vs accepted): 0.455, 95% CI $[0.259, 0.799]$.
- Equivalently: the accepted subset is $1.85\times$ enriched in csPCa+ relative to the rejected subset.

As a hypothesis-generating subgroup signal, this could motivate prospective study of queue-prioritization policies. The rejected subgroup is, on average, approximately half as likely to be csPCa-positive. It is not a validated triage rule. The rejected subgroup still contains csPCa-positive cases ($\frac{15}{95} = 15.8\%$), which is not clinically negligible. Any deployment would require prospective safety constraints. The structural specificity of the cohort effect[4] is supported by the random matched-rate baseline (0.2 ± 5.0 pp shift vs 13.4 pp calibrated; §6.4).

Reading 2, decision-quality null (negative). On the discriminative axis the same subgroup identification *does not* improve clinical decision-making: the three measurements above are flat-to-negative (risk-coverage AUROC unchanged at 0.980; decision-curve net benefit negative, both CIs below zero). The rejected subgroup is structurally lower-risk, but the residual ranking quality on the kept set does not improve. In a workflow where the goal is to make a yes/no diagnostic call on individual patients, the gates do not help.

Why both readings matter for this approach. Reading 1 is the cohort-shift effect; Reading 2 is the detection effect. The framework predicts both. The calibration-tail surface separates a structurally specific sub-distribution (Reading 1) without making the residual target distribution more discriminable (Reading 2). The two are not contradictory. They describe what a target-misaligned surface does and does not deliver. We report both so that readers in different clinical workflows can map the result to their own use case.

6.3 Sub-population analysis: site, age, PSA

We stratified the calibrated Bosma22b rejection rate by center, age, and PSA: the site effect is modest (within a factor of 1.7, overlapping Wilson 95% CIs), age has no detectable effect, and a mild PSA ≥ 10 elevation

[4]The likely driver is the volume/PSA-density confound (Benson et al., 1992): benign prostatic hyperplasia inflates gland volume (tripping the volume gate) while lowering PSA density, so the rejected large-gland subgroup skews benign rather than carrying csPCa signal. We do not stratify by volume here.

is consistent with case-difficulty correlates. None of these overturn the global diagnostic-null reading; full stratification is in Appendix B.

6.4 Baseline comparisons

We benchmark the calibrated rejection against three baselines drawn from existing data:

Rejection set	***N* rej.**	**Containment Δ**	**Sensitivity Δ**	**FP-vol Δ**	**csPCa Δ**
Calibrated p_{10}/p_{90}	95	+0.006	−0.061	0	−13.4 pp
Random matched-rate	95	-0.000 ± 0.051	$+0.004 \pm 0.068$	0	-0.2 ± 5.0 pp
Tightened p_3/p_{97}	251	+0.006	−0.040	0	+3.3 pp

Table 6: Calibrated rejection vs. baselines. All csPCa Δ values use the convention $\Delta_{\text{csPCa}} :=$ rejected rate − accepted rate. Random matched-rate rejection draws 95 cases uniformly at random (1,000 iterations); means and standard deviations reported. Tightened p_3/p_{97} is the same gate logic with stricter percentile envelopes. Random rejection produces a csPCa shift indistinguishable from zero; the calibrated effect (−13.4 pp) is structurally specific. The tightened-threshold variant produces a +3.3 pp csPCa shift in the opposite direction. More aggressive rejection captures more csPCa+ cases, so the rejected set is now slightly enriched, supporting the interpretation that the cohort shift is gate-construction-dependent rather than rejection-rate-dependent. None of the three sets show a containment-fraction or sensitivity gain distinguishable from random.

The baselines support the diagnostic-null reading along two complementary axes. Random rejection isolates how much of the observed cohort shift is structural. With only 0.2 ± 5.0 pp of csPCa shift, vs 13.4 pp calibrated, the calibrated gates are doing *something* that targets a clinically structured subset. Tightening thresholds (which scales rejection rate from 6.3% to 16.7%) reverses the sign of Δ_{csPCa} from −13.4 pp to +3.3 pp on 251 rejected cases. The enlarged rejected set is now slightly csPCa-enriched. This does not flip the diagnostic-null reading: more rejections do not transmute the surface into a csPCa detector (tightened-threshold decision-curve net benefit is not separately measured in this paper).

Section 7.4 separately benchmarks intra-model TTA softmax uncertainty on the same task. It answers a different question: does softmax confidence localize segmentation error? For the theory's central question, does the deployed verification surface separate the target?, the relevant null is random matched-rate rejection on the same surface.

6.5 What the diagnostic-null says about this view

Three observations are worth emphasizing.

1. The framework predicts the negative result: the decision-curve negativity is what one expects when the deployed predicate class is small relative to the target representation.

2. This view provides the remedy direction. Improving csPCa decision quality requires changing the representation, not adding more anatomical gates. A gland-mask representation enriched with tumor-relevant features (lesion-aware geometry, intensity statistics within mask, gland-internal heterogeneity) might support pure-positive buckets for csPCa. The tested tightening sweep did not produce a decision-quality benefit. Read this way, the Reading 1 cohort shift is the readout that selects the move: it is probabilistic enrichment the deterministic chain rule (§2) does not credit, enough to show the target is correlated with what the channel sees but not enough for zero-FP coverage, which points to a new signal-bearing surface rather than more predicates over the same masks. §8.4 and §8.5 construct it.

3. The diagnostic-null protects against overgeneralization. If an evaluator runs the calibrated surface against any clinical target and looks only at cohort effects ($p = 0.0072$ on csPCa+ rate), the surface looks effective. This approach directs the evaluator toward decision-curve and risk-coverage measurements that distinguish cohort from detection. The negative decision-curve result is not failure of the gates. It is a correctly diagnosed surface mismatch.

7 Empirical purity against external targets

Section 6′s null came from one channel, the gland gates. Is it a weakness of those particular gates, or of the gland surface itself? We benchmark the deterministic gates against the two standard medical-imaging QC families, cross-model ensembling (§7.3) and test-time augmentation (§7.4). The comparison is not a competition: each is a distinct verification surface with its own target alignment, and none verifies csPCa at zero false positives, which is what motivates the constructive step in §8. This section measures whether the gates' artifact-level guarantee carries over to other target labels that live at different layers (the strong-reference Dice channel; the patient layer). For targets where the gates are not aligned by construction, both purity and target coverage are empirical. We report the calibrated PI-CAI Stage-2 rejection set's purity $\hat{\rho}$ and target coverage $\hat{\Gamma}$ against two external target families. These are Dice-disagreement shown at three thresholds and csPCa. The operating point is the calibrated p_{10}/p_{90} threshold ($|E_{S_2}| = 95$ rejections):

Target α	$\lvert L_\alpha \rvert$	$\lvert E_{S_2} \cap L_\alpha \rvert$	$\hat{\rho}_{S_2,\alpha}$	$\hat{\Gamma}_{S_2,\alpha}$
Dice < 0.85 (vs Guerbet23)	6	1	1.1% (1/95)	16.7% (1/6)
Dice < 0.90 (vs Guerbet23)	15	3	3.2% (3/95)	20.0% (3/15)
Dice < 0.95 (vs Guerbet23)	87	6	6.3% (6/95)	6.9% (6/87)
csPCa-positive	425	15	15.8% (15/95)	3.5% (15/425)

Table 7: Stage-2 rejection-set purity ($\hat{\rho}$) and target coverage ($\hat{\Gamma}$) against two external target families (Dice-disagreement at three thresholds and csPCa; four rows). None of these targets is aligned with the calibration-tail predicate by construction, so $\hat{\rho} < 1$ in every row. The csPCa+ row's $\hat{\rho} = 15.8\%$ reflects the 13.4 pp cohort shift documented in Section 6. A calibrated rejection lands inside csPCa+ at a lower rate (15.8%) than in the accepted set (29.2% csPCa+).

The pattern is operational: the gates' deterministic-elimination guarantee transfers to the target the gates were designed against (calibration-tail) but degrades sharply on external targets. When a Stage-2 gate fires, the case is certainly outside the calibrated anatomical envelope, but it might or might not be Dice-disagreement, and it is unlikely to be csPCa-positive. No deployed Stage-2 gate or tested Stage-2 rejection union meets the $\rho = 1$ criterion for any of these external targets, so we report $\hat{\rho}$ and $\hat{\Gamma}$ rather than deterministic $\hat{\eta}$. The only zero-false-positive deterministic recall measured in this paper is for the construction-aligned target in Section 5, calibration-tail membership (S_2). A richer predicate class over the same Stage-2 representation might in principle expose a zero-false-positive region for some external target. This view's design ceiling η^* is not bounded by the deployed predicate class we tested here.

7.1 Per-gate ablation

Within Stage 2, only four of eight gates fire on the strong-model PI-CAI Bosma22b dataset. This is informative: gates calibrated on PROMISE12 percentiles produce dormant rejections on a model whose distribution is in-envelope on most predicates. The full per-gate firing table, with per-target breakdowns, is in Appendix C.

The dormancy of four Stage-2 gates on this model is not a defect. It means the deployed pipeline is operating in a regime where those predicates are not stressed. On a weaker model (e.g., TotalSegmentator on prostate, where rejection rate jumps to 93%, Section 9), the dormant gates fire heavily. The same Stage-2 gate construction therefore stays low on strong in-domain models and rises sharply on weaker or off-distribution ones.

7.2 Deterministic baselines: does the Stage-2 gate union outperform single-feature alternatives?

A natural reviewer question is whether the Stage-2 gate union adds anything over its strongest single-gate variant or over a one-feature deterministic baseline. We compare four deterministic, learning-free baselines at or near the calibrated rejection rate (95 cases) on PI-CAI (Bosma22b); the full comparison table is in Appendix D.

Two findings stand out. First, *the Stage-2 gate union is not a single-feature stand-in.* `centroid_smoothness` alone has comparable Dice-disagreement coverage at a smaller rejection set, but `volume_bounds` and `convexity` alone are nearly useless for Dice-disagreement detection. The Stage-2 gate union's value is in target alignment for the calibration-tail target it was designed against, not in being the strongest single Dice-disagreement detector. Second, *the strong-reference baseline is score-aligned with the Dice-disagreement target*: it thresholds the same Dice statistic used to define the target, so coverage is near-tautological. It achieves only $\hat{\Gamma}_{\text{csPCa+}} = 8.0\%$ for csPCa+, however. The calibrated anatomical-tail surface and the strong-reference Dice surface are different surfaces, with different target alignments. Neither dominates the other on every target.

The framework predicts this pattern. Different deterministic baselines align with different targets. The verification-surface analysis identifies *which* predicate construction aligns with *which* target, rather than asserting that one predicate is universally best.

7.3 Split-conformal baseline on model disagreement

Conformal prediction is a standard distribution-free framework for converting any score function into a finite-sample-coverage-controlled rejection rule (Angelopoulos & Bates, 2023; Vovk et al., 2005). To address the natural reviewer question, "how does a learning-free conformal baseline on the same inputs compare?", we implement a split-conformal procedure using $1 - \text{Dice}(\text{Bosma22b}, \text{Guerbet23})$ as the nonconformity score.

Procedure. On a 50/50 calibration/test split of all $N = 1500$ cases (seed $= 42$), the rejection threshold τ is the $(1-\alpha)$-quantile of calibration $1 - \text{Dice}$ scores, used as a selective-rejection rule rather than for marginal label coverage; under exchangeability the realized test rejection rate concentrates on α.

Comparison on the same test split. Because the conformal and TTA baselines are gland-surface comparisons, the deterministic comparator in this subsection is the Stage-2 anatomical-tail gate union, not the full Stage-2 ∪ Stage-4 deployed family. We evaluate both the calibrated Stage-2 gate union and the split-conformal baseline on the held-out test set. Here $n_{\text{test}} = 750$, with $|L_{\text{csPCa+}}| = 207$ and $|L_{\text{Dice}<0.95}| = 37$:

Method	$\lvert E\rvert$	$\hat{\rho}_{\text{Dice}<0.95}$	$\hat{\Gamma}_{\text{Dice}<0.95}$	$\hat{\rho}_{\text{csPCa+}}$	$\hat{\Gamma}_{\text{csPCa+}}$	Notes
Stage-2 gate union (calibrated, on test)	41 (5.5%)	7.3%	8.1%	17.1%	3.4%	Same gates, no recalibration
Conformal (matched-rate, $\alpha = 5.5\%$)	36 (4.8%)	100%	97.3%	47.2%	8.2%	Score-aligned with Dice; uses Guerbet23 as second model
Conformal ($\alpha = 6.3\%$)	37 (4.9%)	100%	100%	46.0%	8.2%	Same observation at slightly higher α
Conformal ($\alpha = 10\%$)	53 (7.1%)	69.8%	100%	41.5%	10.6%	Higher rate, similar enrichment

Table 8: Split-conformal baseline using $1 - \text{Dice}(\text{Bosma22b}, \text{Guerbet23})$ as nonconformity score, evaluated on the held-out test split ($n_{\text{test}} = 750$, seed $= 42$). Conformal achieves higher coverage on the Dice-disagreement target (the score is the target statistic) and higher coverage on csPCa+ ($\hat{\Gamma} = 8.2\%$ vs the gates' 3.4% on this test split). The gates achieve neither by design. They are calibrated for anatomical-tail membership and use only one model.

Reading. Conformal out-performs the gates on both external targets *on this split*, but only by using a richer surface (a second segmenter plus a calibration split). The two are not in competition: they are different surfaces with different inputs, costs, and target alignments, as the framework predicts.

Practical considerations. The conformal baseline re-runs both models per case and needs a calibration split; the gates need only one mask and fixed thresholds, so the cost asymmetry favors the gates.

Decision-curve outcome (csPCa+). We do not extend the §6 decision-curve analysis to conformal; whether its csPCa cohort-enrichment yields decision-curve gain is left to future work.

7.4 Intra-model uncertainty baseline (TTA softmax)

Reviewers of medical-imaging deterministic-QC work commonly ask how the proposed deterministic surface compares to *intra-model* uncertainty. This includes the model's own softmax confidence and its sensitivity to test-time augmentation. We address that question here with a TTA wrapper around the deployed nnU-Net (Isensee et al., 2021) v1 prostate-gland model.

Setup. We use the publicly released nnU-Net v1 prostate-gland model (McCrumb et al., 2023) (BAMF Health Zenodo release, DOI 10.5281/zenodo.8290093[5]) on the $N = 300$ T2W subset of the 1,500-case cohort, split 50/50 into calibration and test ($n_{\text{test}} = 150$, seed = 42). Each case gets a base pass plus three axis-flip passes, softmax-aligned, from which we extract three case-level uncertainty scores:

- **`entropy_mean_fg`**: mean per-voxel Shannon entropy of the base softmax over the predicted-foreground voxels.
- **`tta_dice_mean`**: mean Dice between the base argmax mask and each TTA-augmented argmax mask. Lower means more TTA disagreement.
- **`ensemble_entropy_mean_fg`**: mean entropy of the per-voxel-averaged softmax across the four passes.

Same-subset comparison. As in §7.3, this is a gland-surface comparison, so the deterministic comparator is the Stage-2 anatomical-tail gate union, not the full Stage-2 ∪ Stage-4 deployed family. The §7.3 split-conformal numbers are reported on the full 1,500-case PI-CAI cohort ($n_{\text{test}} = 750$). For a same-denominator three-surface head-to-head, we *recompute* both the gate rejections and the split-conformal rejections on the 150-case TTA test subset. This uses the same seed-42 cal/test partition. All rows in the table below therefore share the same denominator ($n_{\text{test}} = 150$).

Results. On the same-subset 150-case test split ($|L_{\text{Dice}<0.95}| = 12$, $|L_{\text{csPCa+}}| = 48$):

Method	$\lvert\boldsymbol{E}\rvert$	$\hat{\boldsymbol{\rho}}_{\textbf{Dice}<\textbf{0.95}}$	$\hat{\boldsymbol{\Gamma}}_{\textbf{Dice}<\textbf{0.95}}$	$\hat{\boldsymbol{\rho}}_{\textbf{csPCa+}}$	$\hat{\boldsymbol{\Gamma}}_{\textbf{csPCa+}}$	**Comment**
Stage-2 anatomical-tail gate union	14 (9.3%)	14.3%	16.7%	28.6%	8.3%	Reference: deterministic gates
TTA `entropy_mean_fg`	13 (8.7%)	0.0%	0.0%	23.1%	6.3%	Single-fold base softmax entropy on FG
TTA `tta_dice_mean`	15 (10.0%)	0.0%	0.0%	33.3%	10.4%	Per-case Dice variance under flips
TTA `ensemble_entropy_mean_fg`	13 (8.7%)	0.0%	0.0%	23.1%	6.3%	Entropy of TTA-averaged softmax
Conformal ($\alpha = 6.3\%$, on 150-case subset)	13 (8.7%)	92.3%	100%	46.2%	12.5%	Cross-model Bosma/Guerbet Dice score

Table 9: Three-surface comparison on the 150-case TTA test subset (seed = 42). All five rows share the same denominator. On the strong-reference Dice-disagreement target, conformal-on-disagreement dominates by construction because it uses the Dice score directly. Anatomical-tail gates catch some cases (16.7%) while intra-model TTA catches none. On csPCa+ enrichment, conformal-on-disagreement is again strongest, with `tta_dice_mean` second and gates third. No single surface dominates across targets. The §7.3 split-conformal numbers on the full 750-case test split ($\hat{\Gamma} = 100\%$ on Dice < 0.95, $\hat{\Gamma} = 8.2\%$ on csPCa+) are consistent with these subset numbers within sampling noise.

Where the surfaces disagree: pairwise Jaccard overlap. If "different surfaces, different alignments" is the right reading, the three surfaces should reject *disjoint* sets of cases. They do:

[5]This is the public-release nnU-Net v1 prostate-gland model trained on PI-CAI data, distributed as a single nnU-Net model export (`Task788_Prostate.zip`). It is the same nnU-Net v1 architectural family and the same prostate-gland T2 MRI task as the original Bosma22b challenge submission. The weights are from the BAMF Health release, not Bosma22b's algorithm container, and we do not claim shared training-data lineage. The intra-model uncertainty baseline is therefore representative of the nnU-Net v1 prostate-gland surface. Bit-exact reproduction of Bosma22b's ranking would require its algorithm container.

Surface pair	$\lvert A \cap B \rvert$	$\lvert A \cup B \rvert$	Jaccard
Gates ∩ Conformal	2	33	0.061
Gates ∩ TTA `entropy_mean_fg`	2	25	0.080
Gates ∩ TTA `tta_dice_mean`	0	29	0.000
Conformal ∩ TTA `entropy_mean_fg`	0	34	0.000
Conformal ∩ TTA `tta_dice_mean`	0	36	0.000
TTA `entropy_mean_fg` ∩ TTA `tta_dice_mean`	10	18	0.556

Table 10: Pairwise Jaccard overlap of rejection sets on the 150-case test split (matched rejection rate $\approx 9\%$). Cross-surface overlap is near zero ($J \leq 0.08$). Within-family overlap (the two TTA scores derived from the same softmax) is $J = 0.556$, as expected. The three verification surfaces, anatomical-tail gates, cross-model disagreement, and intra-model TTA, reject almost completely different cases at the same rejection rate. This makes the view's prediction directly visible at the case-set level.

Reading. TTA does reject cases on this split, but *different* ones (mostly csPCa-negative), not the Dice-disagreement cases the anatomical-tail gates catch.

Practical considerations. TTA adds roughly 4 × per-case GPU cost where the gates add none, and shows no compelling compute/benefit tradeoff on this subset.

7.5 Reading §7 as a surface×target alignment map

Consolidating the §7.3 and §7.4 results into a single surface×target view makes the framework's diagnostic prediction empirically visible from data already in the paper. Each row is a verification surface (a representation plus a predicate class); each column is a target.

Surface	Calibration-tail	Dice < 0.95	csPCa+
Stage-2 anatomical-tail gates	aligned ($\rho = \eta = 1$, §5)	$\hat{\rho} = 14.3\%$, $\hat{\Gamma} = 16.7\%$	$\hat{\rho} = 28.6\%$, $\hat{\Gamma} = 8.3\%$
Cross-model disagreement (conformal)	not aligned	score-aligned ($\hat{\rho} = 92.3\%$, $\hat{\Gamma} = 100\%$)	$\hat{\rho} = 46.2\%$, $\hat{\Gamma} = 12.5\%$
Intra-model uncertainty (best TTA)	not aligned	$\hat{\rho} = 0\%$, $\hat{\Gamma} = 0\%$	$\hat{\rho} = 33.3\%$, $\hat{\Gamma} = 10.4\%$

Table 11: Surface×target alignment map for the three verification surfaces compared in §7.3 and §7.4. All Dice-and-csPCa numbers come from the same 150-case test split. The "not aligned" cells for conformal and TTA on calibration-tail reflect representation mismatch (neither score encodes the calibrated anatomical envelope).

Three structural facts the map exposes. *First*, each surface has at most one target it is naturally aligned with. Gates align with calibration-tail by construction; conformal aligns with Dice disagreement because its nonconformity score IS the target statistic; TTA aligns with none of the displayed targets. *Second*, no surface dominates across targets, and the §7.4 pairwise Jaccard ≤ 0.08 confirms this at the rejection-set level. *Third*, against csPCa+, none of the three surfaces reaches alignment; every cell shows weak-to-moderate enrichment rather than $\hat{\rho} = 1$. This is the framework's diagnostic-value claim made empirically visible on data already in the paper: the three baseline comparator surfaces evaluated in §7, binary-mask gates, cross-model disagreement, and intra-model uncertainty, do not carry sufficient signal to achieve useful zero-FP deterministic coverage of csPCa+ cases. The csPCa diagnostic null is therefore not a property of any one chosen channel; it is a property of all three comparator surfaces evaluated here, predicted by the representation-product structure of the targets they expose. To raise alignment on csPCa+, the natural-next-contribution direction (§12) is to add a representation that carries the cancer signal, not to engineer better predicates over these three surfaces. §8.4–§8.5 reports the Phase-2 pilot in that direction, instrumenting a fourth surface, a deliberately engineered ADC-based product predicate, that achieves zero observed test false positives on the same held-out split where the three §7 baselines do not.

8 Cross-stage complementarity

Sections 5–7 stayed on a single surface and found that no single surface verifies csPCa. We now ask what *composition* buys: stacking artifact-level gates from two different verification surfaces, the Stage-2 gland-mask surface and the Stage-4 lesion-list surface, in series. The architectural prediction is that pipeline miss probabilities shrink multiplicatively when stage rejection events catch largely different subsets of target-positive cases. We test this with three measurements: stage-level rejection-set Jaccard cross-overlap ($\omega_{\text{stage-cross}}$), conditional escape ratio against two main target labels (with a third broad lesion-containment sensitivity target reported in Appendix E), and greedy cross-stage portfolio coverage relative to same-stage portfolios.

8.1 Stage-level rejection cross-overlap $\omega_{\text{stage-cross}}$

Pooled across the three domains, the stage-level Jaccard overlap of rejection sets is small.

Domain	$\lvert E_{S_2}\rvert$	$\lvert E_{S_4}\rvert$	$\lvert E_{S_2} \cap E_{S_4}\rvert$	**ω Jaccard**	**Bootstrap 95% CI**
Prostate	95	50	2	1.40%	[0.00%, 3.68%]
Liver	0	8	0	0.00%	[0.00%, 0.00%]
Kidney	46	5	1	2.00%	[0.00%, 6.67%]
Pooled	**141**	**63**	**3**	**1.49%**	[0.51%,4.30%] Wilson

Table 12: Stage-level rejection-set Jaccard overlap. Pooled Wilson 95% CI on $\frac{3}{201}$ is $[0.51\%, 4.30\%]$; pooled bootstrap (case-resampling) 95% CI is $[0.00\%, 3.37\%]$. Liver has zero Stage-2 rejections under the nnU-Net model used in this comparison; its Fisher test is degenerate with $p = 1.0$. Pooled Cochran–Mantel–Haenszel test gives common odds ratio $= 0.81$ ($p = 0.73$).

For all three domains, observed stage-level co-rejection is not inconsistent with independent rejection margins at the rejection-set level. The prostate Fisher test has $p = 0.77$, kidney has $p = 0.39$, and the pooled CMH test has $p = 0.73$. (This is a statement about co-rejection, not about escape independence; the latter is reported separately in §8.2.) Per-stage gate density (combined-family Jaccard within Stage 2) is much higher: 0.183 (95% CI $[0.133, 0.230]$). The low-overlap pattern is visible in prostate and kidney, while liver contributes a dormant-stage boundary case ($|E_{S_2}| = 0$). The result is also stable across the tested Stage-4 representations.

8.2 Conditional escape factorization

Write $U_{S_k,\alpha}$ for the event that a target-positive case escapes the gates at stage k (the per-stage escape event $\left(E_{k,\alpha}\right)^c$ of §2). The *marginal escape* $P\left(U_{S_4}\right)$ is the rate at which Stage 4 misses overall; the *conditional escape* $P\left(U_{S_4} \mid U_{S_2}\right)$ is the rate at which it misses among the cases Stage 2 also missed.

Across two tested target labels α, we measure $P\left(U_{S_4} \mid U_{S_2}\right) - P\left(U_{S_4}\right)$ and the ratio $\frac{P\left(U_{S_4} \mid U_{S_2}\right)}{P\left(U_{S_4}\right)}$. The chain rule (Section 2) is the population identity; the question is whether the conditional and marginal miss probabilities are equal in the sample.

Target α	$\lvert L_\alpha\rvert$	$P\left(U_{S_2,\alpha}\right)$	$P\left(U_{S_4,\alpha}\right)$	$P\left(U_{S_4} \mid U_{S_2}\right) - P\left(U_{S_4}\right)$	**Ratio (95% CI)**
Low gland Dice < 0.95	87	93.1%	94.3%	+2.0 pp $[-0.3, +5.5]$	$1.022[0.997, 1.060]$
csPCa-positive	425	96.5%	88.2%	+0.1 pp $[-0.5, +0.8]$	$1.001[0.995, 1.009]$

Table 13: Conditional escape relative to marginal escape. Direct rates use Wilson 95% CIs; ratios and differences use 10,000-iteration case-bootstrap CIs. Both ratio CIs contain unity; both difference CIs contain zero. Conditional-escape diagnostics are numerically close to the independent-product approximation at this sample resolution, with no claim of strict independence.

For both targets the ratio's 95% CI contains 1.0 and the difference's 95% CI contains 0.0. The multiplicative decomposition is consistent with the data at the resolution of this sample. The Fisher exact test on the

low-Dice target gives $p = 0.0364$, the only marginally significant result of the two; it is driven by a 2.0-pp residual that the bootstrap CI nonetheless cannot distinguish from zero. For the csPCa target, we detect no departure from the product-like conditional-escape approximation (Fisher $p = 0.69$). Appendix E reports the same diagnostic for a broad lesion-containment-disagreement sensitivity target.

8.3 Stage-coverage beats same-stage density (rejection coverage)

The framework's stage-coverage corollary predicts that, at a matched gate budget b (the number of gates in the portfolio), a portfolio that spreads gates across stages should outperform a portfolio concentrated at a single stage. The natural test is *rejection-set coverage*: starting from $b = 1$, each step adds the gate that catches the most new rejections. (Section 10 reports the analogous test against external *target-label coverage* on cross-domain tumor-present targets, where the result is target-dependent.)

b	Best overall	Best same-stage	Δ	95% CI (Δ); CI > 0?
1	3.67% (55)	3.67% (s1)	0.00 pp	[0.00, 0.00] pp; no
2	6.80% (102)	6.00% (s1)	+0.80 pp	[0.00, 1.93] pp; no
3	9.13% (137)	6.27% (s1)	+2.87 pp	[2.00, 3.80] pp; **yes**
4	9.40% (141)	6.33% (s1)	+3.07 pp	[2.20, 4.00] pp; **yes**
5	9.47% (142)	6.33% (s1)	+3.13 pp	[2.33, 4.07] pp; **yes**

Table 14: Greedy cross-stage gate portfolios vs. best same-stage portfolios at matched b. From $b = 3$ onward, cross-stage portfolios dominate same-stage portfolios with bootstrap 95% CIs excluding zero. Random portfolios at $b = 5$ achieve mean coverage $4.90\% \pm 2.31\%$, well below the 9.47% greedy cross-stage figure, ruling out a "more gates = more rejections" trivial explanation.

At $b \geq 3$, the cross-stage delta exceeds 2 pp with bootstrap 95% CIs that exclude zero, *for rejection-set coverage*. A random-portfolio baseline (1,000 random subsets per b, seed = 42) caps at $4.90\% \pm 2.31\%$ at $b = 5$. That is far below the greedy cross-stage figure, ruling out a trivial "more gates = more rejections" explanation. The result on external target-label coverage is more nuanced. Kidney tumor-present shows a small significant cross-stage gain, while liver tumor-present shows no gain because S_2 is dormant in that domain. Section 10 reports those results.

8.4 Phase 2: ADC as a complementary intensity-aware surface

The paper so far has characterized the deliberately chosen mask-only channel: exact alignment with the construction-aligned calibration-tail target (§5), diagnostic null on the clinical csPCa target (§6), and cross-stage composition within the mask-only channel (§8.1–§8.3). That mask-only composition enriches the csPCa cohort, raising prevalence among flagged cases from 15.8% to 44.1%, but stays decision-curve-negative (§6): it amplifies enrichment without reaching a target the channel does not represent. The next two subsections use the framework constructively: given the csPCa diagnostic-null result, what surface shape should be added? The §6 diagnostic null fixes the form of the answer: not another predicate over the same masks, but a new verification surface over a representation that exposes the target, composed with the existing one. Trust Topology predicts that adding a representation containing csPCa-relevant signal should produce new target-positive buckets. ADC provides a minimal empirical test of that prediction.

This subsection reports a sensitivity analysis using a single ADC-derived intensity surface. The §2 representation-product composition mechanism predicts that two representations over the same case space, with different representation maps Y_A and Y_B, form a joint representation that refines each marginal. We test this empirically with a minimal intensity-aware fourth surface alongside the binary-mask gates. The comparator throughout is the Stage-2 anatomical-tail gate union, not the full Stage-2 ∪ Stage-4 deployed family: ADC is a gland-support surface, so the gland channel is the like-for-like comparator, and the Stage-4 lesion gates depend on the Bosma22a lesion mask, whose csPCa-label training would contaminate the very complementarity claim under test.

The predicate is a single ADC scalar, $s = q_{05}(\text{ADC[gland]}) \,/\, \text{median}(\text{ADC[gland]})$, computed inside the Bosma22b gland mask. The threshold τ_s was calibrated to a matched rejection rate (the closest calibration-split rejection rate not exceeding the Stage-2 anatomical-tail gate-union rate on the locked 152-case test split), giving $\tau_s = 0.4960$, $10/148 = 6.8\%$ calibration rejections, and $13/152 = 8.6\%$ test rejections. This matched-rate choice uses the Stage-2 rejection-set size only and does not use test csPCa labels; it is a descriptive operating-point comparison on the locked test split, not a zero-FP validation claim.

At matched rejection rate, the ADC and Stage-2 rejection sets were disjoint on the held-out test split: gates rejected 11 cases, ADC rejected 13, the intersection was 0, the union was 24, and rejection-set Jaccard was 0.000.

Target	**Test T+**	**Gates**	**ADC**	**Both (∩)**	**Either (∪)**
csPCa+	44	2	7	0	9
Dice < 0.95	13	3	2	0	5
ISUP ≥ 3	20	1	4	0	5
Stage-2 anatomical-tail rejection	11	11	0	0	11

Table 15: Matched-rate ADC complementarity on the held-out test split. Counts are target-positive catches at a matched 7% rejection rate; the "Both (∩)" column is the intersection of the two rejection sets and "Either (∪)" their marginal union. None of these cells is a formal alignment claim or an $\hat{\rho} = 1$ claim. The Stage-2 anatomical-tail target row reads the redundancy axis directly: ADC catches 0 of the cases the gates would catch, confirming that ADC is not detecting "what the gates would catch" but a different class.

The csPCa+ row is the headline. Using two marginal predicates each calibrated to similar rejection rates ($\approx$ 7%), the union catches 9/44 (20.5%) csPCa+ test cases where Stage-2 gates alone catch 2/44 (4.5%), a 4.5 $\times$ lift in observed csPCa+ coverage. The union rejection rate is $24/152 = 15.8\%$, the sum of the two marginal rates because the surfaces' firing sets are disjoint (Jaccard = 0.000). The csPCa+ escape set shrinks from 42 cases that escape gates-only $(44 - 2)$ to 35 cases that escape the marginal union $(44 - 9)$. ISUP ≥ 3 shows a similar shape ($1/20 \to 5/20$, 5 $\times$ lift on the high-grade subset). Neither surface aligns with csPCa+ or ISUP ≥ 3 at $\rho = 1$; the lift is observed target-positive coverage from unioning two disjoint rejection sets, not guaranteed zero-FP deterministic capacity.

What this demonstrates and what it does not. This analysis demonstrates the marginal-union *lower bound* on representation-product composition: rejecting when either marginal fires expands observed csPCa+ coverage from 4.5% (gates alone) to 20.5% (union) at a 15.8% combined rejection rate. What it does not demonstrate is joint refinement, a predicate over both surfaces' values jointly identifying pure-positive buckets neither marginal recognizes; that is §8.5.

8.5 Above-marginal-union refinement: a clinically-grounded predicate

§8.4 established the marginal-union lower bound; we now construct a predicate that achieves the above-marginal-union refinement §2 predicts. To find such a predicate without building a learned csPCa model or soliciting bespoke expert-designed rules, we surveyed the prostate mpMRI literature for published deterministic operating points and selected three candidate predicate forms before any test evaluation.

Engineering selection rule. Candidate Phase-2 surfaces were screened against four framework-derived requirements: (i) expose an information channel absent from the mask-only predicates; (ii) remain deterministic and auditable; (iii) use a representation whose support artifact is not learned from csPCa labels; and (iv) be testable as a product predicate over the same case space so that above-marginal-union refinement could be measured. Prostate MRI literature then supplies the domain prior: restricted diffusion on ADC is a known csPCa-relevant signal. C1 is the simplest candidate satisfying all four requirements. C2 and C3 satisfy (i), (ii), and (iv) but fail (iii) because they use a Bosma22a lesion mask as their support artifact, and Bosma22a is trained on csPCa labels. They are reported in Appendix F as caveated lesion-mask channel comparators: their published thresholds confirm that the ADC operating points catch csPCa+ cases on

this cohort, but their support inherits target signal, so they cannot bear the headline above-marginal-union refinement claim.

The simplest candidate with csPCa-label-separated support combines a PI-RADS-range ADC threshold with a structural connected-component requirement inside the Bosma22b gland mask:

$$P_{\mathrm{C1}}(\text{case}) = [\text{largest 6-connected component volume of} \\ \{x \in \text{gland} : \mathrm{ADC}(x) < 800 \times 10^{-6}\mathrm{mm}^2/\mathrm{s}\} \\ > 4.44 \text{ cc}]$$

The ADC threshold range comes from PI-RADS v2.1 (Radiology, 2019) (markedly restricted diffusion at < 750–$900 \times 10^{-6}\mathrm{mm}^2/\mathrm{s}$); a histopathology-validated $< 682 \times 10^{-6}\mathrm{mm}^2/\mathrm{s}$ mean-ADC cutoff for peripheral-zone csPCa is reported in Costa et al. (Costa et al., 2019). The operating thresholds ($\tau_{\mathrm{ADC}} = 800$, $\tau_{\mathrm{vol}} = 4.44$ cc) are selected on the calibration split for zero observed false positives on csPCa+, with the volume constrained to ≥ 0.5 cc to avoid singleton-voxel components.

On calibration: 5/148 firings, 5/43 csPCa+ catches, 0 FP. Held-out test, single application with frozen thresholds: 6/152 firings, 6/44 csPCa+ catches ($\hat{\Gamma} = 13.6\%$), 0 FP. This is a finite-sample held-out $\hat{\rho} = 1$ candidate operating point on a 152-case test split, not a population-level zero-FP guarantee. Wilson 95% bounds quantify the small-sample uncertainty: test FPR $\leq 3.4\%$ given 0/108 csPCa− firings, $\hat{\Gamma} \in [6.4\%, 26.7\%]$ from 6/44, and above-marginal-union coverage $4/44 \in [3.6\%, 21.2\%]$.

Composition with §8.4 surfaces on csPCa+:

Surface	**Test firings**	**csPCa+ caught**	**Rejection-set Jaccard with C1**
Stage-2 anatomical-tail gates	11	2	0.000 (0/17)
§8.4 matched-rate ADC marginal	13	7	0.118 (2/17)
§8.4 marginal union (gates ∪ ADC)	24	9	0.071 (2/28)
C1 (this section)	6	6	—
Combined: gates ∪ §8.4 ADC ∪ C1	28	13	—

Table 16: Surface stacking on csPCa+ at the operating points calibrated in their respective sections. C1′s rejection set is disjoint from the Stage-2 gates' rejection set and overlaps the §8.4 ADC marginal's rejection set on 2 cases (both csPCa-positive: $10540_{1000551}$, $10956_{1000975}$). C1 contributes 4 csPCa+ catches above the §8.4 marginal union: cases $10522_{1000532}$, $10699_{1000715}$, $11341_{1001364}$, $11352_{1001375}$. The combined stack catches 13/44 (29.5%) csPCa+ test cases at $28/152 = 18.4\%$ rejection rate. The combined stack is *not* zero-FP on csPCa+: Stage-2 gates contribute 9 csPCa-negative firings, the §8.4 ADC marginal contributes 6, and these are disjoint from each other and from C1, giving 15 csPCa-negative firings against 13 csPCa-positive catches. C1 itself has zero observed test FP. The combined stack increases observed target-positive coverage by unioning C1 with lower-purity surfaces.

Purity and coverage rise together. Intuitively, bolting on another test should produce *more* false flags, not fewer: a new gate rejects more cases, the extra rejections mix true hits with false alarms, and precision drifts down toward the base rate. That is the ordinary precision-coverage tradeoff, and it is what unioning surfaces usually buys. Adding C1 does the reverse. On csPCa+ the operating-point ladder reads: Stage-2 gates alone, purity $2/11 = 18\%$ at coverage 4.5%; gates-plus-ADC marginal union, $9/24 = 37\%$ at 20.5%; adding C1, $13/28 = 46\%$ at 29.5%. The reason it inverts is exact: C1′s four added rejections are *all* true positives (4/4 csPCa+, zero observed FP), so coverage rises with no new false alarms and purity climbs instead of decaying. The framework gives the condition (a unioned-in surface raises purity only when its novel rejections are purer than the current pool); C1, at zero observed FP, is the limiting case. The bare marginal union does not invert: folding the low-purity gates into the higher-purity ADC marginal is the ordinary trade.

This both-rise is §2′s refinement monotonicity made empirical, and it is a refinement effect, not union superadditivity: a union catches at most the sum of its disjoint parts, so the four above-union catches are not the union exceeding arithmetic but the joint (mask, ADC) representation creating pure-positive buckets

neither marginal recognizes. The theorem makes the effect a predicted possibility, not a fluke; the four catches are a finite-sample realization of it (purity $\hat{\rho} = 1$ on the 152-case test, not a certified population capacity), which is why this stays a constructive pilot. And it requires the new representation: no composition within the mask channel reaches this regime, exactly as the §6 diagnostic null predicts.

Secondary target ISUP ≥ 3 shows the same pattern. With $\tau_{\text{vol}} = 5.10$ cc recalibrated for the ISUP target, C1 catches 3/20 ($\hat{\Gamma} = 15\%$) at zero observed test FP, with 2/3 outside the §8.4 union.

Provenance. C1 uses the Bosma22b gland mask and raw ADC voxel values. Bosma22b is not trained on csPCa labels, but it is trained on PI-CAI gland-segmentation data, so the support is not cohort-external. The provenance claim is therefore label-channel separation, not full cohort independence: the representation is csPCa-label-untrained, while the operating point is csPCa-calibrated. Scanner protocol, cohort composition, and PI-CAI acquisition style may still be reflected in the learned gland mask. External validation should therefore use a pre-specified C1 predicate on an independent cohort or a gland support generated without PI-CAI training.

Other candidates evaluated. Two additional literature-derived candidates were evaluated with frozen thresholds and are reported in the released supplemental analysis files; both were excluded from the headline composition table because they depend on Bosma22a lesion-mask region selection and Bosma22a is trained on csPCa labels, weakening provenance for this surface-characterization claim.

Methodological note: standing on the shoulders of giants. The point is not that Trust Topology discovers ADC. Radiologists and the prostate MRI literature already identify diffusion restriction as clinically relevant. The framework's contribution is to say *why* ADC must enter as a new verification surface rather than as more threshold tuning over binary masks, and *how* to test whether the resulting product surface adds coverage beyond the marginals. The predicate form was not invented here: the ADC threshold range and the focal-lesion-size criterion both come from clinical literature accumulated against histopathology truth over decades of prostate MRI research.[6] §2 tells us in advance what kind of predicate to look for (a deterministic function over a representation that exposes a previously absent information channel), and §6 tells us why the naive predicate fails (csPCa is out-of-regime for the mask-only channel alone). The literature provides the predicate shape and clinically plausible threshold ranges. The calibration split selects the operating point. The framework provides the criterion that turns clinical knowledge into a verification surface.

9 Same gates, four models

Sections 5–8 held the upstream model fixed (Bosma22b). The same gate construction, applied to four models of differing quality, drives rejection from 6.3% to 93% with zero gate changes.

Within the prostate domain, the same Stage-2 gland gates and fixed prostate thresholds evaluate four PI-CAI gland-segmentation models of widely differing quality. This is a Stage-2 gland-model comparison. The Stage-4 lesion gates are not part of it, and cross-domain liver/kidney are excluded since their thresholds are domain-recalibrated:

[6]The literature survey used an LLM-assisted deep-research workflow to compile candidate predicate forms. The workflow generated a menu of approximately 20 (target, predicate) pairs from prostate mpMRI guidelines and primary literature; we down-selected three candidates that satisfied the engineering requirements (i), (ii), and (iv) above and reported all three in Appendix F. The workflow was a search aid, not evidence; all retained predicate shapes and thresholds are traced to the cited peer-reviewed sources.

Model	Quality (gland Dice)	Stage-2 rejection rate	Notes
Bosma22b	≈ 0.90 (strong)	6.3% ($\frac{95}{1500}$)	Reference model used elsewhere in this paper
Guerbet23	≈ 0.88 (strong)	$\approx 5\%$	Reference model for Dice disagreement
MONAI	≈ 0.85 (medium)	$\approx 11\%$	General-purpose pretrained
TotalSegmentator	≈ 0.45–0.80 (weak; out-of-domain)	93%	General CT model applied off-distribution

Table 17: Same Stage-2 gates, four models. Rejection rate stays low for strong in-domain models and rises sharply for the off-distribution model, from 6.3% on Bosma22b to 93% on TotalSegmentator.

The architecture-level reading of this table is the theory's:

- Same gates. Same thresholds. Different models.
- The pipeline does not have to be redesigned for each model. The verification surface is model-independent at the gate-construction level; only the upstream model output changes.

This is the "verification amplification" property: the same surface gives more rejections when the upstream model is worse, without any change to the surface itself. The operational consequence is that the surface is a reusable reliability asset: engineered once against the target, it carries over to new or upgraded models without redesign, so reliability tracks the surface rather than the model behind it.

The same surface doubles as a label-free model-quality benchmark. The rejection rate in the table is computed from the masks alone, with no ground-truth segmentations, yet it tracks the quality column: low for the strong in-domain models, rising sharply as the model becomes a poor fit for the domain, to 93% on off-distribution TotalSegmentator. The construction that rejects bad artifacts is therefore also a way to *rank* upstream models, using none of the expert labels the gland-Dice column requires. The scalar rate reports how good a model is; the structure of its firings reports how it fails, since on a strong model the rejections stay near-disjoint across stages (the §8 and §10 complementarity result, $\omega = 1.40\%$) while a model that is a poor domain fit saturates the surface.

10 Cross-domain transfer

Everything so far has been prostate. This view predicts that the architectural rule, cross-stage complementarity (§8), should hold whenever the per-stage representations expose anatomical predicates that produce nearly disjoint target sets. Testing this prediction requires the rule to transfer across organs, modalities, and models. We replicate the gate-family construction on liver tumor segmentation (MSD Task03, CT, $N = 131$, 11 gates). We also replicate it on kidney tumor segmentation (KiTS23, CT, $N = 489$, 9 gates after disabling two anatomically inapplicable predicates). Stage 2 becomes organ segmentation. Stage 4 becomes tumor segmentation. Thresholds are recalibrated within domain (no PI-CAI thresholds used cross-domain).

10.1 Stage-level Jaccard overlap by domain

Domain	ω	Wilson 95% CI	Notes
Prostate (Bosma22b/Bosma22a)	1.40% ($\frac{2}{143}$)	$[0.39\%, 4.95\%]$	Reference baseline
Liver (nnU-Net)	0.00% ($\frac{0}{8}$)	$[0.00\%, 32.4\%]$	Stage-2 rejection rate is zero under nnU-Net
Kidney (nnU-Net CV)	2.00% ($\frac{1}{50}$)	$[0.35\%, 10.50\%]$	Cross-validated nnU-Net predictions
Pooled	**1.49%** ($\frac{3}{201}$)	$\mathbf{[0.51\%, 4.30\%]}$	Wilson on pooled $\frac{3}{201}$

Table 18: Cross-domain stage-level Jaccard overlap. Pooled 1.49% across three domains. Prostate and kidney show low cross-stage overlap with nonzero Stage-2 firing. Liver contributes a dormant-stage boundary case with zero Stage-2 rejections ($|E_{S_2}| = 0$) under this model and threshold, so its overlap is trivially zero rather than an independent replication. The complementarity finding holds across organ, modality, and model for the two domains with active Stage-2 gates.

10.2 Cross-domain target-label coverage

Greedy gate-portfolio coverage against tumor-present targets, by domain. *Interpretation note.* The KiTS23 evaluation subset used here is tumor-positive throughout ($N = 489$, all tumor+). The kidney row therefore measures *target-positive coverage* (which gates catch tumor-positive cases) rather than tumor-vs-non-tumor *discrimination*. The MSD liver subset has 118 tumor-positive cases (out of $N = 131$). The 13 tumor-negative cases are present but not used in the tumor-present recall computation. Discrimination claims would require additional tumor-negative cohorts.

Domain	$\lvert L_\alpha \rvert$	Any-gate recall	Cross-stage delta	Significant b
Liver tumor present	118	3.4% ($\frac{4}{118}$)	0.0 pp	none
Kidney tumor present	489	10.2% ($\frac{50}{489}$)	+0.8 pp at $b = 4, 5$	$b = 4, 5$ (CIs > 0)

Table 19: Greedy cross-stage portfolios against tumor-present targets. Kidney tumor presence yields a small but positive cross-stage delta at $b = 4, 5$ (bootstrap CI excludes zero); liver tumor presence has zero delta at all b, consistent with the small liver Stage-2 rejection set.

The kidney result supports the architectural rule against an external target-positive coverage label, tumor presence, though this is not tumor-vs-non-tumor discrimination given the all-tumor-positive KiTS23 subset used here. Cross-stage portfolios of size $b = 4, 5$ outperform same-stage portfolios with bootstrap 95% CIs excluding zero. The liver pipeline at this model and threshold has zero Stage-2 rejections, leaving no room for cross-stage gain.

Taken together, the three replications make the structural claim domain-general: cross-stage complementarity is a property of pipeline composition, not of prostate anatomy. The same gate-family construction, recalibrated per domain, reproduces the near-disjoint rejection structure across organ, modality, and model; only the upstream artifacts and thresholds change, not the surface logic.

11 Discussion

11.1 What this means for the framework

The flagship's coding-agent study (Rothrock, 2026a) develops the verification surfaces on a revision-architecture system (prompt → plan → design → code → test). In that system, artifacts are generated transformations and refinement-monotonicity is an *operational analogue*, not a literal mathematical refinement. The medical pipeline is a non-revision architecture: each case flows through Stage 1 → Stage 2 → Stage 3 → Stage 4 once. The same operational pattern applies. The Stage-2 representation introduces a mask-geometry surface that exposes anatomical-tail predicates not available as simple predicates on raw pixels. Because the segmentation model is a learned (and stochastic) transformation, this is not a guaranteed

measure-theoretic refinement of the raw image. It is an operational refinement that empirically exposes predicates the upstream representation does not.

The cross-stage chain rule is more directly testable in a non-revision architecture, where each case has a single Stage-2 artifact and a single Stage-4 artifact. Section 8 reports the empirical content of that test.

Quality control for medical imaging segmentation has historically been treated as either an uncertainty-estimation problem (Bayesian or ensemble methods) or an out-of-distribution-detection problem (calibrated softmax, energy-based detectors). Both approaches are stochastic by construction: they produce a continuous score and a tunable threshold.

The verification-surface view is complementary to both, with a different aim. The framework asks an operational question: for which targets can the gates *deterministically eliminate* cases from the accepted output space, meaning that when a gate fires, the case is certainly inside the declared violation event, or equivalently outside the declared acceptable operating envelope, without a probabilistic estimate? Some targets admit this guarantee (calibration-tail, by construction); others do not (csPCa, because the channel does not contain the signal). The theory's diagnostic value is distinguishing the two *before* engineering effort goes into improving thresholds. Formally: targets where $\rho = 1$ is achievable over the chosen predicate class.

Composition, and how patient-level claims emerge from artifact-level guarantees. Each gate is a deterministic statement about one artifact. Stacked, the gates form a chain: if any one fires, the case is held for review. The reach of that chain is bounded by which artifacts are in it. The mask-only gates in Phase 1 of this paper cover the Stage-2 gland-mask surface and the Stage-4 lesion-list surface. The mask channel alone (without intensity-aware extensions) may carry weak correlates of clinical status (the §6 cohort shift is one such correlate), but it does not expose enough of the pathology and mpMRI signal channels to support a validated patient-level csPCa claim. The constructive direction is to compose more informative artifacts into the chain. Adding gates over the Stage-1 raw-image artifact (T2W intensity, ADC values, gland-internal heterogeneity) would let the chain reach signal channels that include patient-relevant information. Phase 2 of this paper (§8.4–§8.5) takes the first pilot step in that direction with a single ADC-derived intensity surface, demonstrating both marginal-union and above-marginal-union refinement on csPCa+ at finite-sample held-out scale. The composed surface does not become a patient-level csPCa detector; it spans channels that move a patient-level test closer to admissibility. Extending the chain with broader mpMRI-aware artifacts (see §12, *The natural next contribution*) is the next link.

Operational consequences of crossing the composition threshold. Four practical consequences follow from these composition results (§8.1–§8.3 cross-stage coverage, §8.5 representation-product purity), and they jointly account for why composition (not single-channel verification) is the operationally relevant unit. Each is a candidate for prospective evaluation rather than a deployment claim. *First*, in a prospective workflow that handled the predicate-defined artifact failures covered by the composed surface deterministically, the human reviewer's residual responsibility would narrow. The reviewer would no longer be the catch-all for the implausible-mask-volume, disconnected-component, or lesion-outside-gland classes that this channel does declare; the reviewer's vigilance would be reserved for failure modes this mask-only channel explicitly does not represent. *Second*, composition can raise the purity of the flagged subset when an added surface is aligned to the target (§8.5 demonstrates this on the clinical target), turning the flag into a plausible prioritization signal. The radiologist is not relieved of their primary read, but a higher-purity flag is sharp enough to motivate prospective evaluation as a directed second-look signal rather than decay into ignorable noise. *Third*, the deterministic predicate-level coverage that survives composition protects downstream automation that has no human in the loop. Automated lesion-volume measurement, treatment-planning calculations, cohort selection for clinical trials, and model-monitoring pipelines all consume artifact-level masks without per-case human review. Each additional composed channel widens the predicate-level coverage that protects those downstream automated steps from silently inheriting artifact-level deviations the channel can verify; the surface's deterministic accept event *is* the precondition that makes the downstream automation auditable against this set of predicates. *Fourth*, every accept and reject in the composed surface is traceable to a specific predicate violation. This is the audit trail subjective human review does not produce. When a clinical decision based on a broken AI output is later investigated, the gate record shows whether

the composed surface accepted the case under its declared predicates; if it did, the failure mode is outside the declared predicate set (or at least outside the predicates currently deployed over this channel) and becomes a candidate for predicate expansion or a new verification surface. A single channel can provide a local version of each of these benefits. Composition makes them operationally stronger by widening the predicate-defined failure classes covered before downstream use.

On learned QC classifiers. Learned QC heads, separately trained models that take segmentation outputs as input and predict quality, are not benchmarked here. They are complementary to the verification-surface analysis, not strictly above or below it. A learned QC head can in principle improve any of the three surfaces compared in Section 7 by adding a trained classifier on top of the deployed predicate. This view predicts that such a head would still be constrained by the underlying surface's target alignment. A head trained on calibration-tail labels would help the gates' surface. A head trained on Dice-disagreement labels would help the conformal surface. *Which target the head is trained against is itself a verification-surface design choice.* Empirical evaluation of learned QC heads is left to follow-on work.

The reliability question at the heart of this paper is not specific to prostate MRI. Multi-stage AI pipelines in any domain share the structural property that quality assurance happens at intermediate artifact boundaries. Examples include coding agents (the flagship's pipeline), legal document review, autonomous-driving perception, radiology triage, and scientific simulation post-processing. In each case, target alignment determines what each boundary can verify. The medical-imaging case is unusually clean (discrete artifacts, public expert labels, deterministic predicates), which is why we use it as the empirical anchor. The transferable claim is the framework, not the gates. *Verification surfaces are target-specific; no single uncertainty or QC mechanism is universally dominant; reliability engineering for multi-stage AI begins with surface-target alignment, not with global model trust.*

This view reframes a familiar engineering choice. When a multi-stage pipeline has a target it cannot serve well, the standard responses are to retrain the model, add data, tune thresholds, or change architectures. This approach adds a fourth option: change the representation at the verification surface. Section 6 shows this concretely. Tightening anatomical thresholds did not improve csPCa decision quality in this experiment. The framework predicts that further improvement requires a representation exposing csPCa-relevant structure rather than merely stricter thresholds over the same anatomical-tail surface. The remedy is a different surface, gland mask + intensity + geometry features that expose csPCa-relevant predicates, not better thresholds on the existing one.

12 Limitations

The limits below describe the scope and boundaries of the Phase-1 mask-only characterization and the Phase-2 constructive pilot. Each names a place where extending the channel, the predicate class, or the cohort is the natural next contribution.

Negative-control rejection sets are a robust check against confusing cohort with detection. The 13.4 pp csPCa-rate shift on calibrated rejection is real, in the sense that random matched-rate rejection produces only 0.2 ± 5.0 pp. But the calibrated effect, while structurally distinguishable from random, is still a cohort shift, not detection. Decision-curve analysis distinguishes the two cleanly. Risk-coverage analysis with cross-validated calibration provides an independent second check. Both yield the same conclusion: the deployed gates show no decision-quality benefit for csPCa beyond reshaping the cohort.

Single-pipeline empirical scope. The empirical results report one prostate pipeline (PI-CAI) with cross-domain replication on liver and kidney. While the architectural rule transfers across domains, the pipeline-specific findings (model-quality scaling on TotalSegmentator, decision-curve negativity for csPCa) are properties of these specific pipelines. Generalization to other prostate datasets, other csPCa scoring methods, or other clinical outcomes would require domain-specific replication.

Predicate-class restrictions. The deployed gates are simple anatomical predicates with calibrated thresholds. The Surface Capacity characterization characterizes the design ceiling under arbitrary measurable

predicates over the same representation, not under this restricted class. The csPCa diagnostic-null in Section 6 is therefore a result about the deployed surface, not a proof that no deterministic predicate over gland masks can separate csPCa.

Calibration-set size and transferability. Stage-2 thresholds are calibrated on $N = 50$ PROMISE12 expert segmentations. The transfer to PI-CAI is non-trivially good for the strong/in-domain models. Rejection rates are well-controlled for Bosma22b, Guerbet23, and MONAI on the 1,500-case and 300-case cohorts respectively. TotalSegmentator's 93% rejection rate is the intentional off-distribution case (§9) and is not described as well-controlled. Tail-percentile estimates from 50 calibration cases still have wide confidence intervals. Cross-domain liver and kidney recalibration is even more constrained (no cross-domain calibration set comparable to PROMISE12).

Conditional escape sample sizes. Section 8′s conditional escape ratios on the two main target labels have CIs that contain unity, but the small-$|L_\alpha|$ low-Dice case has wide ratio CIs ($[0.997, 1.060]$, $|L_\alpha| = 87$). Larger samples, especially for targets defined by strong-reference disagreement, would tighten the test of the product-like miss-reduction approximation.

Expert-label provenance. All ground-truth comparisons use the expert annotations released with the source competitions. These are population-level expert labels, not sampled audits, so individual-rater agreement is not estimated here. Where the original challenges report inter-annotator variability (PI-CAI, KiTS23), that variability bounds the achievable precision of any predicate evaluated against a single expert label.

Phase-2 is constructive screening, not clinical validation. The ADC/C1 extension is included to demonstrate the constructive engineering move predicted by the framework: after a mask-only diagnostic null, add a representation that exposes a clinically relevant information channel and test product predicates over the joint surface. It is not presented as a validated csPCa detector. Three literature-derived candidate forms were evaluated on the same 152-case test split after calibration-split threshold selection, so the C1 6/6 zero-observed-FP result has candidate-selection exposure and should be read as a promising candidate operating point, not a pre-registered single-candidate validation. The above-marginal-union refinement count itself is finite-sample: 4/44 csPCa+ above-union catches with Wilson 95% CI $[3.6\%, 21.2\%]$, so any single missing case would shift the headline meaningfully. C1 also uses Bosma22b gland support, which is csPCa-label-untrained but PI-CAI-cohort-trained; the provenance claim is label-channel separation, not full external independence. A confirmatory test should pre-specify the predicate, use an independent external cohort or independently generated gland support, and evaluate csPCa coverage, purity, ISUP stratification, and decision-curve net benefit once, without candidate selection on the test cohort.

Framework scope and falsifiability. Trust Topology does not, by itself, identify ADC as the correct prostate-cancer feature; clinical literature supplies that domain prior. The framework contributes the engineering criterion: if a target is not exposed by the current representation, more predicates over that same representation should not be expected to produce zero-FP deterministic coverage; the next constructive move is to expose a representation carrying the missing signal and test deterministic predicates over the resulting product surface. The falsifiable units in this paper are therefore surface-target claims, not the broad slogan. The mask-only diagnostic-null claim would be weakened by a deterministic predicate over binary segmentation masks alone (no intensity inputs) that achieves test FP $= 0$ on csPCa+ with $\hat{\Gamma} > 5\%$ on a separate cohort. The cross-stage composition claim would be weakened by pooled cross-stage Jaccard > 0.20 or by no marginal coverage gain (within bootstrap CI) from greedy cross-stage portfolios over same-stage portfolios at $b \geq 3$. The Phase-2 constructive claim would be weakened if a pre-registered, comparably powered family of intensity-aware deterministic product predicates, calibrated on a separate cohort, produced no above-marginal-union csPCa+ catches at zero observed false positives on an external held-out cohort.

The natural next contribution. The Phase 2 exploratory extension in §8.4–§8.5 is the first pilot step in this direction, not the completed clinical validation. The natural next contribution is a pre-specified, adequately powered mpMRI-aware verification surface that combines T2W intensity statistics, DWI/ADC features inside both gland and lesion regions, gland-internal heterogeneity, and lesion-morphology descrip-

tors, with thresholds fixed before external validation and evaluated against clinical endpoints (csPCa+, ISUP grade, lesion-volume, and decision-curve net benefit on a held-out clinical cohort). PI-CAI releases all input modalities publicly; the experiment is direct.

13 Conclusion

This pipeline demonstrates the three regimes of verification-surface engineering in a single empirical setting: aligned artifact targets can be eliminated exactly; stage surfaces compose when their rejection sets are disjoint; clinical targets outside the channel require a new representation. The ADC pilot shows the constructive next step: adding a target-relevant surface creates coverage beyond the mask-only marginals, while remaining a finite-sample candidate rather than a validated detector.

The framework is demonstrated here on prostate segmentation and transfers to liver and kidney, supporting the reading that its structural claims are about pipeline composition, not prostate anatomy. Together with the flagship and the language companion, this work positions verification surfaces, not models, as the unit of reliability for multi-stage AI pipelines.

Appendix A Stage-4 `lesion_volume` calibration audit

The `lesion_volume` envelope (≤ 50 cc) was loosened from an initial literature-informed 30 cc ceiling against this cohort's expert lesion annotations, the within-cohort calibration disclosed in §3. Two facts bound its effect. First, the adjustment is immaterial to the focal-lesion retention claim: every expert mask in the 30–50 cc band is itself a confluent, whole-gland-scale annotation (a single component occupying 40–72% of the gland), so the original 30 cc ceiling would not have falsely rejected any valid focal lesion either. The looser bound only widens the accept region, so any residual within-cohort bias is toward under-rejection, not toward inflating the gate's apparent performance. Second, the gate is nearly inert: across the 1,500-case cohort the volume and count envelopes reject one case (an over-segmented mask exceeding the gland on every axis), and on a held-out split (seed 42, $n_{\text{test}} = 750$) they falsely reject zero expert-positive test cases, so the adjusted bound generalizes. On the full 425-case expert-positive population the volume envelope rejects two expert masks; both are single connected components occupying 52–83% of the gland and spanning nearly every gland slice (one with marksheet evidence of two ISUP 5 lesions). These are confluent, whole-gland-scale annotations rather than focal lesions, correctly flagged as implausible, with the 50 cc bound sitting in a sparse tail (99th percentile of expert lesion volume = 40.5 cc).

Appendix B Subpopulation stratification (site, age, PSA)

PI-CAI cases come from three centers (RUMC, PCNN, ZGT). We stratify the calibrated Bosma22b rejection rate by center, patient age, and PSA. The main-body reading (§6) is that none of these effects overturns the diagnostic-null conclusion.

Stratum	n	Rejected	Rate	Wilson 95% CI
Center: RUMC	800	42	5.2%	[3.9%, 7.0%]
Center: PCNN	350	22	6.3%	[4.2%, 9.3%]
Center: ZGT	350	31	8.9%	[6.3%, 12.3%]
Age $<$ 60	284	18	6.3%	[4.0%, 9.7%]
Age 60–69	770	47	6.1%	[4.6%, 8.0%]
Age $\geq$ 70	446	30	6.7%	[4.7%, 9.4%]
PSA $<$ 4	132	7	5.3%	[2.6%, 10.5%]
PSA 4–10	761	39	5.1%	[3.8%, 6.9%]
PSA $\geq$ 10	567	48	8.5%	[6.5%, 11.0%]

Table 20: Sub-population stratification of Stage-2 calibrated rejection rate (Bosma22b, $|E_{S_2}| = 95$). Center effect: ZGT (8.9%) vs RUMC (5.2%), modest site variation, with overlapping CIs at the 95% level. Age: stable across buckets (no age effect). PSA: similar for < 4 and 4–10 subgroups, slightly elevated for PSA ≥ 10 (8.5%). The csPCa+ rate among PSA ≥ 10 patients is also elevated, so the gates may interact mildly with PSA-correlated case difficulty. We treat this as a calibration-tail correlate, not a confounder, and the diagnostic-null reading on csPCa+ holds across PSA strata.

Appendix C Stage-2 per-gate ablation

Within Stage 2, only four of eight gates fire on the strong-model PI-CAI Bosma22b dataset (§7). This table is the per-gate evidence behind the stage-coverage corollary (§8.3): on a strong model, packing more gates into one stage adds little coverage, so the leverage is cross-stage composition rather than denser gating within a stage. The full per-gate firing, with per-target breakdowns, is below.

Gate	$\|E_g\|$	**csPCa+ caught**	**Dice < 0.95 caught**	$\hat{\rho}_{\text{csPCa+}}$	**Comment**
`centroid_smoothness`	55	14	5	25.5%	Strongest single gate; carries most of the firing
`slice_area`	35	1	0	2.9%	Catches none of the strong-reference disagreements
`volume_bounds`	33	1	0	3.0%	Volume alone is uninformative for Dice-disagreement
`convexity`	2	1	2	50.0%	Tiny rejection set; high purity by chance
`single_component`	0	—	—	—	Dormant on Bosma22b
`anatomical_location`	0	—	—	—	Dormant on Bosma22b
`slice_continuity`	0	—	—	—	Dormant on Bosma22b
`aspect_ratio`	0	—	—	—	Dormant on Bosma22b

Table 21: Stage-2 per-gate firing on PI-CAI Bosma22b, with per-target breakdowns. Four of eight gates are dormant on this model, supporting this approach's stage-coverage corollary. At fixed gate budget, gates that catch the same cases (or no cases) are not worth deploying. What matters is whether each gate adds rejection coverage. The greedy ablation in §8.3 makes this concrete: by $b = 5$, the optimal portfolio adds at most one new case beyond $b = 4$, indicating diminishing per-gate contribution within Stage 2 alone.

Appendix D Deterministic single-feature baselines

We compare four deterministic, learning-free baselines at or near the calibrated rejection rate (95 cases) on PI-CAI (Bosma22b): the strongest single gate (`centroid_smoothness`), a volume-only threshold, a second-model disagreement score, and a random control. The full per-gate breakdown, including the four dormant gates, is in Appendix C. The main-body conclusion (§7) is that the Stage-2 union is not a single-feature stand-in, and that different deterministic surfaces align with different targets.

Method	$\lvert E\rvert$	$\hat{\rho}_{\text{Dice<0.95}}$	$\hat{\Gamma}_{\text{Dice<0.95}}$	$\hat{\rho}_{\text{csPCa+}}$	$\hat{\Gamma}_{\text{csPCa+}}$	**Notes**
Stage-2 calibrated anatomical-tail union	95	6.3%	6.9%	15.8%	3.5%	Reference (Stage-2 gates only; not the full Stage-2+Stage-4 family)
Single gate: `centroid_smoothness`	55	9.1%	5.7%	25.5%	3.3%	Strongest single gate
Volume-only p_{10}/p_{90}	300	3.0%	10.3%	33.0%	23.3%	3× rejection rate; lower purity
Strong-reference disagreement ($\text{Dice}_{\text{BvG}} < 0.955$, matched rate)	94	92.6%	100%	36.2%	8.0%	Score-aligned with Dice target; uses second model
Random matched-rate	95	$s\,\text{im}\,5.8\%$ *	$s\,\text{im}\,6.3\%$ *	28.3% *	6.3% *	Mean over 1,000 draws, seed $= 42$ (Dice prevalence $\frac{87}{1500} \approx 5.8\%$; matched-rate coverage expectation $\frac{95}{1500} \approx 6.3\%$)

Table 22: Deterministic baselines compared on PI-CAI (Bosma22b). The single-gate variant (`centroid_smoothness`) reduces the deployed family to one predicate. The strong-reference baseline rejects when Bosma22b vs Guerbet23 Dice falls below a threshold matched to the calibrated 95-case rejection rate. This is *score-aligned* with the Dice-disagreement target (the predicate thresholds the same Dice statistic at a slightly looser cutoff, so it catches every Dice < 0.95 case and $\hat{\Gamma} = 100\%$ by threshold inclusion) but instructive against csPCa+. Random matched-rate values marked with * are simulation means over 1,000 random draws of the same rejection rate (seed = 42). The calibrated effect on csPCa-rate is 13.4 pp vs random's 0.2 ± 5.0 pp.

Appendix E Sensitivity target: broad lesion-containment disagreement

As a sensitivity check on the cross-stage independence finding (§8.2), we run the same conditional-escape diagnostic against a broad lesion-containment-disagreement target: the 284 cases whose Bosma22a lesion mask places more than 5% of its volume outside the Bosma22b gland mask. This is a target, not a gate. The deployed Stage-4 `lesion_in_gland` gate fires only above 50% outside, so most of this broader target escapes it; the gate is deliberately conservative, rejecting gross containment failure rather than ordinary boundary disagreement between the two mask sources.

Target α	$\lvert L_\alpha\rvert$	$P\big(U_{S_2,\alpha}\big)$	$P\big(U_{S_4,\alpha}\big)$	$P\big(U_{S_4} \mid U_{S_2}\big) - P\big(U_{S_4}\big)$	**Ratio (95% CI)**
Lesion-containment disagreement $> 5\%$	284	97.2%	82.4%	+0.2 pp $[-0.5, +1.2]$	$1.003[0.994, 1.014]$

Table 23: Conditional-escape diagnostic for the broad lesion-containment-disagreement sensitivity target. The deployed Stage-4 gate (at $> 50\%$ outside) catches 50 of the 284 cases in this broader $> 5\%$ target, which is why Stage-4 escape is high (82.4%).

The ratio's 95% CI contains 1.0 and the difference's CI contains 0.0 (Fisher $p = 0.63$), consistent with the same product-like conditional-escape pattern at this sample resolution reported for the low-Dice and csPCa targets in §8.2.

Appendix F Phase-2 candidate screen

To make the §8.5 multiple-testing exposure visible, this appendix tabulates the three literature-derived Phase-2 candidates. All three were frozen as forms before any test evaluation, calibrated for zero observed false positives on the calibration split and applied once to the locked 152-case test split. C1 is the §8.5 headline.

Candidate	Representation	Threshold source	Test firings	csPCa+ firings	Observed FP	Status
C1	Bosma22b gland + ADC scalar + connected component	PI-RADS v2.1 range; calibration	6	6	0	headline candidate (§8.5)
C2	Bosma22a lesion + mean ADC	Costa et al. histopathology cutoff	1	1	0	excluded: csPCa-trained support
C3	Bosma22a lesion + ADC + HBV ratio + diameter	PI-RADS v2.1 DWI 5 conjunction	1	1	0	excluded: csPCa-trained support

Table 24: Phase-2 candidate screen on the locked seed-42 test split ($n_{\text{test}} = 152$, csPCa+ test total = 44). All three candidates passed the calibration-split zero-observed-FP rule; all three produced zero observed test FP. C2 and C3 were excluded from the headline composition table because Bosma22a is trained on csPCa labels, so its lesion mask is not csPCa-label-separated support. C1 is reported as a finite-sample candidate operating point with candidate-selection exposure, not a pre-registered single-candidate validation.

References

Angelopoulos, A. N., & Bates, S. (2023). Conformal Prediction: A Gentle Introduction. *Foundations and Trends in Machine Learning*, *16*(4), 494–591. https://doi.org/10.1561/2200000101

Antonelli, M., Reinke, A., & Bakas, S. (2022). The Medical Segmentation Decathlon. *Nature Communications*, *13*, 4128. http://medicaldecathlon.com/

Benson, M. C., Whang, I. S., Pantuck, A., Ring, K., Kaplan, S. A., Olsson, C. A., & Cooner, W. H. (1992). Prostate Specific Antigen Density: A Means of Distinguishing Benign Prostatic Hypertrophy and Prostate Cancer. *The Journal of Urology*, *147*(3 Part 2), 815–816.

Bosma, J. S., Alves, N., & Huisman, H. (2022,). *Performant and Reproducible Deep Learning-Based Cancer Detection Models for Medical Imaging*. Annual Meeting of the Radiological Society of North America. https://fastmri.eu/research/bosma22b.html

Bosma, J. S., Saha, A., Hosseinzadeh, M., Slootweg, I., Rooij, M. de, & Huisman, H. (2023). Semisupervised Learning with Report-Guided Pseudo Labels for Deep Learning-Based Prostate Cancer Detection Using Biparametric MRI. *Radiology: Artificial Intelligence*, *5*(5), e230031. https://fastmri.eu/research/bosma22a.html

Cardoso, M. J., Li, W., Brown, R., Ma, N., Kerfoot, E., & Wang, Y. (2022). MONAI: An Open-Source Framework for Deep Learning in Healthcare. *arXiv preprint arXiv:2211.02701*. https://arxiv.org/abs/2211.02701

Cochran, W. G. (1954). Some Methods for Strengthening the Common Chi-Square Tests. *Biometrics*, *10*(4), 417–451.

Costa, D. N., Xi, Y., Aziz, M., Passoni, N., Shakir, N., Goldberg, K., Francis, F., Roehrborn, C. G., Leon, A. Diaz de, & Pedrosa, I. (2019). Prospective Inclusion of Apparent Diffusion Coefficients in Multiparametric Prostate MRI Structured Reports: Discrimination of Clinically Insignificant and Significant Cancers. *American Journal of Roentgenology*, *212*(1), 109–116. https://doi.org/10.2214/AJR.18.19937

Debs, N., Routier, A., Abi-Nader, C., Marcoux, A., Nicolas, F., Bone, A., & Rohe, M.-M. (2023,). *DataScientX Algorithm Trained on PI-CAI: Private and Public Training Dataset*. https://grand-challenge.org/algorithms/pi-cai-pubpriv-datascientx/

DIAGNijmegen. (2025, July 1). *Annotations for the PI-CAI Public Training and Development Dataset* (No. commit c6bab0b (tag v2.0-37-gc6bab0b). GitHub. https://github.com/DIAGNijmegen/picai_labels

Fisher, R. A. (1935). *The Design of Experiments*. Oliver and Boyd.

Heller, N., Isensee, F., & Trofimova, D. (2023). *The KiTS21 Challenge: Automatic Segmentation of Kidneys, Renal Tumors, and Renal Cysts in Corticomedullary-Phase CT*. https://arxiv.org/abs/2307.01984

Isensee, F., Jaeger, P. F., Kohl, S. A. A., Petersen, J., & Maier-Hein, K. H. (2021). nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation. *Nature Methods*, *18*(2), 203–211. https://doi.org/10.1038/s41592-020-01008-z

Litjens, G. (2014). Evaluation of prostate segmentation algorithms for MRI: The PROMISE12 challenge. *Medical Image Analysis*, *18*(2), 359–373. https://promise12.grand-challenge.org/

Mantel, N., & Haenszel, W. (1959). Statistical Aspects of the Analysis of Data from Retrospective Studies of Disease. *Journal of the National Cancer Institute*, *22*(4), 719–748.

McCrumb, D., Murugesan, G. K., Soni, R., & Van Oss, J. (2023). *Pretrained model for 3D semantic image segmentation of the prostate from T2 MRI scans*. Zenodo. https://doi.org/10.5281/zenodo.8290093

Pooch, E. H. P., Agrotis, G., & Cai, L. (2025). Semi-Supervised Learning in Prostate MRI Tumor Segmentation Approaches Fully-Supervised Performance on External Validation. *Medrxiv*. https://www.medrxiv.org/content/early/2025/05/13/2025.05.13.25327456

Radiology. (2019). *PI-RADS v2.1: Prostate Imaging - Reporting and Data System, Version 2.1*. American College of Radiology. https://www.acr.org/-/media/ACR/Files/RADS/Pi-RADS/PIRADS-V2-1.pdf

Rothrock, M. (2026b). *Interpretable Cross-Stage Quality Control for AI Medical Imaging Pipelines*. Zenodo. https://doi.org/10.5281/zenodo.19362420

Rothrock, M. (2026a). *Trust Topology: Verification Surfaces as the Unit of Reliability*. Zenodo. https://doi.org/10.5281/zenodo.20292194

Saha, A., Bosma, J. S., & Twilt, J. J. (2024). Artificial Intelligence and Radiologists in Prostate Cancer Detection on MRI (PI-CAI): An International, Paired, Non-Inferiority, Confirmatory Study. *The Lancet Oncology*, *25*(7), 879–887. https://pi-cai.grand-challenge.org/

Vickers, A. J., & Elkin, E. B. (2006). Decision Curve Analysis: A Novel Method for Evaluating Prediction Models. *Medical Decision Making*, *26*(6), 565–574.

Vovk, V., Gammerman, A., & Shafer, G. (2005). *Algorithmic Learning in a Random World* (1st ed.). Springer. https://doi.org/10.1007/b106715

Wasserthal, J., Breit, H.-C., Meyer, M. T., Pradella, M., Hinck, D., Sauter, A. W., Heye, T., Boll, D. T., Cyriac, J., Yang, S., Bach, M., & Segeroth, M. (2023). TotalSegmentator: Robust Segmentation of 104 Anatomic Structures in CT Images. *Radiology: Artificial Intelligence*, *5*(5), e230024. https://pubmed.ncbi.nlm.nih.gov/37795137/

Wilson, E. B. (1927). Probable Inference, the Law of Succession, and Statistical Inference. *Journal of the American Statistical Association*, *22*(158), 209–212.

Yates, F. (1934). Contingency Tables Involving Small Numbers and the Chi-Square Test. *Supplement to the Journal of the Royal Statistical Society*, *1*(2), 217–235.

Version 1.0 · July 2026.