Interpretable Cross-Stage Quality Control for AI Medical Imaging Pipelines

Michael Rothrock
Independent Researcher
Correspondence to michael@roth.rocks

Abstract

Multi-stage AI pipelines for medical image analysis can produce structurally implausible outputs (disconnected segmentation masks, lesions placed entirely outside the organ) that propagate to downstream analysis and scoring. We evaluate whether simple, deterministic quality control checks at multiple pipeline stages can catch such failures in a model-agnostic, interpretable, and low-cost manner.

We implemented deterministic quality gates, pure functions requiring no GPU or learned parameters, at two stages of a prostate cancer detection pipeline: organ segmentation (8 gates) and lesion detection (3 gates). Gates encode anatomical plausibility constraints ranging from hard structural rules (e.g., the gland must be a single connected component) to empirical bounds (e.g., organ volume within calibrated range). Thresholds were calibrated on 50 PROMISE12 [8] expert segmentations and applied without modification to 1,500 PI-CAI [10] cases across four segmentation models of widely different quality.

Gate rejection rates scaled with model weakness: 4.8% (multi-center nnU-Net ensemble), 6.3% (prostate-specific nnU-Net), 11% (MONAI prostate model), 93% (TotalSegmentator multi-organ model), all using identical gate code and thresholds. The central finding was cross-stage complementarity: Stage 2 (organ) and Stage 4 (lesion) gates caught largely non-overlapping failures, with only 2 shared cases out of 143 total rejections for Bosma22b. Gate-rejected cases showed degraded downstream lesion detection metrics (Mann-Whitney, $p < 0.05$ on three independent tests across models), and gate filtering improved lesion containment more than random case removal for two of three models (bootstrap test, 10,000 iterations).

Cross-domain validation on liver tumor segmentation (CT, N=131) and kidney tumor segmentation (CT, N=489) replicated the cross-stage complementarity pattern with 0–2% overlap in all three domains. The gate interface and composition logic are universal, while gate selection and thresholds are domain-specific.

These results suggest that deterministic QC gates, approximately 200 lines of Python per domain, can provide interpretable, auditable failure detection across AI imaging pipeline stages. They complement rather than replace model improvement or uncertainty estimation.

Keywords: quality control, AI pipeline reliability, medical image segmentation, deterministic verification, cross-domain validation, clinical deployment

1. Introduction

AI models for prostate cancer detection achieve strong benchmark performance, but clinical deployment exposes a gap between average-case metrics and worst-case reliability. A segmentation model with excellent mean Dice can still produce outputs with disconnected components, erratic centroid trajectories, or implausible volumes. These failures violate basic anatomical constraints and can propagate through multi-stage pipelines, degrading downstream lesion detection and clinical decision support.

Model improvement reduces the frequency of such failures but does not eliminate them, because it targets average performance metrics rather than worst-case anatomical plausibility. A complementary approach is to add explicit quality control checks at each pipeline stage: deterministic functions that flag outputs violating pre-specified anatomical constraints before they reach the next processing step. This approach assumes that the clinically meaningful intermediate artifacts of the process (gland masks, lesion masks) should remain explicit and individually verifiable rather than being absorbed entirely into an end-to-end stochastic predictor.

This idea is not novel in principle; radiologists routinely reject segmentations that "don't look right," and QA programs for imaging hardware use deterministic checks against known standards. What has received less attention is the **composition** of such checks across pipeline stages: how much additional failure coverage does a second stage of verification provide, and are the catches complementary or largely redundant?

We investigate this question using prostate cancer detection from multiparametric MRI as a testbed. We implement 11 deterministic quality gates at two pipeline stages (gland segmentation and lesion detection) and evaluate them across four segmentation models of widely different quality on 1,500 PI-CAI cases. Our main contributions are:

1. **Model-agnostic transfer.** Gates calibrated on 50 expert segmentations transfer without modification across four models on 1,500 cases from a separate dataset, confirming that gates test output properties rather than model-specific behaviors.
2. **Cross-stage complementarity.** Gates at different pipeline stages catch nearly disjoint failure sets (96% of lesion-stage catches were not detected by any gland-stage gate), suggesting that multi-stage QC provides substantial coverage beyond what any single-stage system can achieve.
3. **Selective failure detection.** Gate filtering improves downstream lesion containment more than random case removal, indicating enrichment for cases with downstream-relevant structural inconsistencies rather than merely unusual outputs.

2. Related Work

2.1 Quality Assurance in Medical Imaging

QA for imaging hardware (the ACR's phantom-based accreditation, automated scanner calibration protocols) provides a well-established model for deterministic verification against known standards. Our work extends this model from hardware to AI pipeline outputs: just as a phantom scan verifies that a scanner meets spatial resolution requirements, a quality gate verifies that a segmentation output meets anatomical plausibility requirements.

2.2 Uncertainty Estimation

Uncertainty estimation methods, including Bayesian dropout [1], deep ensembles [2], and test-time augmentation [3], provide estimates that flag unreliable predictions. These stochastic methods address a different failure mode than deterministic gates. Uncertainty estimation catches boundary imprecision and distributional shift; gates catch gross structural violations (impossible volume, disconnected components, spatial inconsistency). The two approaches are complementary. This paper focuses on the deterministic component to isolate its contribution.

2.3 Segmentation Quality Estimation

Post-hoc quality metrics (Dice, Hausdorff distance) require ground truth unavailable at inference time. Surrogate quality estimation without ground truth has been explored using shape priors, atlas comparison, and learned quality scores [4–6]. Our gates differ in requiring no learned parameters, providing binary accept/reject decisions with interpretable reasons, and being trivially composable across pipeline stages.

---

3. Methods

3.1 Gate Design

A quality gate is a deterministic function $g: output \rightarrow \{accept, reject\}$ that checks an anatomical plausibility constraint. Gates fall into two categories:

Hard structural constraints test conditions treated here as structural consistency requirements for anatomically valid output. A prostate gland is a single connected organ; a segmentation with two disconnected components is structurally inconsistent regardless of clinical context. For these gates, a rejection indicates a strong structural abnormality.

Empirical plausibility bounds test conditions derived from observed anatomical ranges. A prostate volume of 250 cc is implausible based on clinical populations, but not logically impossible; rare cases at the tail of the distribution exist. For these gates, a rejection indicates high-confidence implausibility, not certainty.

Both categories share the same operational property: they are pure functions, require no GPU or learned parameters, and produce a human-readable rejection reason. The distinction matters for interpretation: a hard-constraint rejection (disconnected mask) warrants different clinical handling than a plausibility-bound rejection (large but potentially real volume).

Overlap ratio (ω). For a system of n gates with rejection sets $E_1, \ldots, E_n$, we measure redundancy as:

$$\omega = 1 - |E_1 \cup \ldots \cup E_n| / (|E_1| + \ldots + |E_n|)$$

ω = 0 indicates perfectly complementary gates (each catches unique failures); ω = 1 indicates perfect redundancy. This metric quantifies the marginal value of adding gates to an existing system.

3.2 Pipeline Architecture

The prostate cancer detection pipeline processes multiparametric MRI through five stages:

Stage	Function	Output
1	Image quality assessment	Valid/invalid sequences
2	**Gland segmentation**	Prostate boundary mask
3	Zone segmentation	Peripheral/transition zone
4	**Lesion detection**	Suspicious region masks
5	Clinical decision	PI-RADS score, biopsy recommendation

This paper applies gates at **Stage 2** (8 gates) and **Stage 4** (3 gates). Stages 1, 3, and 5 remain ungated.

3.3 Stage 2: Gland Segmentation Gates

Eight gates check plausibility constraints for prostate gland segmentation. Thresholds were calibrated on 50 PROMISE12 [8] expert segmentations, achieving 49/50 pass rate (the single failure is a known annotation outlier including seminal vesicles at 315 cc).

Gate	Type	Condition	Threshold
single_component	Hard	Connected components	Exactly 1
slice_continuity	Hard	Axial gaps in mask	0 gaps
volume_bounds	Empirical	Total mask volume	10–150 cc
convexity	Empirical	Volume / convex hull	≥ 0.80
anatomical_location	Empirical	Centroid position in FOV	Inner 70%
centroid_smoothness	Empirical	Max inter-slice centroid shift	≤ 20% of max diameter

Gate	Type	Condition	Threshold
aspect_ratio	Empirical	Dimensional ratios	1:4 to 4:1
slice_area	Empirical	Max axial slice area	≤ 35 cm²

3.4 Stage 4: Lesion Detection Gates

Three gates check plausibility constraints for lesion detection output. These operate on the lesion mask with access to the gland mask from Stage 2.

Gate	Type	Condition	Threshold
lesion_in_gland	Empirical	Fraction of lesion inside gland	≥ 50%
lesion_volume	Empirical	Total lesion volume	0.01–50 cc
lesion_count	Empirical	Connected components	≤ 10

Thresholds were validated against 220 expert lesion annotations from PI-CAI. No expert cases violated the volume or count thresholds in the validation set. The containment gate flagged 19 expert cases below the 50% threshold: these are cases where the AI gland mask does not fully cover the expert lesion, indicating gland-lesion spatial inconsistency (a legitimate QC finding, though the root cause is upstream gland segmentation error rather than lesion misplacement).

3.5 Analysis Cohorts

Different analyses use different subsets of the data. To avoid confusion, we specify each cohort explicitly:

Cohort	N	Used for
Full PI-CAI	1,500	Gate generalization, clinical correlation, cross-stage composition (Bosma22b and Guerbet23)
Common 300	300	MONAI and TotalSegmentator comparison (cases with imaging available)
Expert lesion (full)	220 of 1,500	Cascading failure analysis (Bosma22b), verification amplification (Bosma22b full)
Expert lesion (common)	47 of 300	Cascading failure and amplification (MONAI, TotalSegmentator)

The 300-case common set is a subset of the full 1,500, limited by availability of imaging data for MONAI and TotalSegmentator inference. Expert lesion annotations exist for 220 of 1,500 PI-CAI cases (all clinically significant cancer cases with expert delineations); 47 of these fall within the 300-case common set.

3.6 Models Evaluated

Four segmentation models of different quality, applied to PI-CAI with identical gate code and thresholds:

Model	Architecture	Domain	Expected Quality
Guerbet23	nnU-Net 5-fold ensemble [11, 12]	Prostate-specific (multi-center)	Strong+
Bosma22b	nnU-Net [7, 11]	Prostate-specific	Strong
MONAI prostate_mri_anatomy	UNet	Prostate-specific	Medium (~Dice 0.85)
TotalSegmentator [9]	Multi-organ	104-structure CT	Weak for prostate

Guerbet23 is a five-fold nnU-Net ensemble trained on the full PI-CAI dataset (1,500 cases from 11 centers, 7 scanners) and was used to provide prostate segmentations for the PI-CAI challenge [10, 12]. It represents a strong, modern, task-matched baseline, the key test case for whether gates still provide value when segmentation quality is high.

TotalSegmentator is not designed for prostate segmentation and the 93% rejection rate reflects this domain mismatch. We include it not as a fair benchmark but as a stress test: if gates are truly model-agnostic, they should activate proportionally to the degree of anatomical implausibility in the output, regardless of why the output is poor.

3.7 Statistical Methods

Mann-Whitney U test (one-sided): Compares downstream metric distributions between gate-accepted and gate-rejected groups, with effect size $r = 1 - 2U/(n_1 \cdot n_2)$.

Fisher's exact test: Tests association between gate decisions and clinical outcomes (csPCa prevalence).

Bootstrap verification amplification test: For each gate configuration, we perform 10,000 iterations of random case removal (removing the same number of cases as gates reject) and compute the downstream metric improvement for each iteration. If actual gate-filtered improvement exceeds the 95th percentile of the random-removal distribution, the gates are selecting failures non-randomly with respect to that metric.

3.8 Operational Handling of Rejected Cases

In a clinical deployment, a gate rejection would route the case for manual review rather than silently passing it through the pipeline. The rejection reason (e.g., "gland volume 263 cc exceeds 150 cc bound" or "47% of lesion voxels outside gland mask") provides the reviewing clinician with a specific,

interpretable flag. Alternative operational responses (fallback to a secondary model, request for re-acquisition, or abstention from automated scoring) are feasible but not evaluated in this paper.

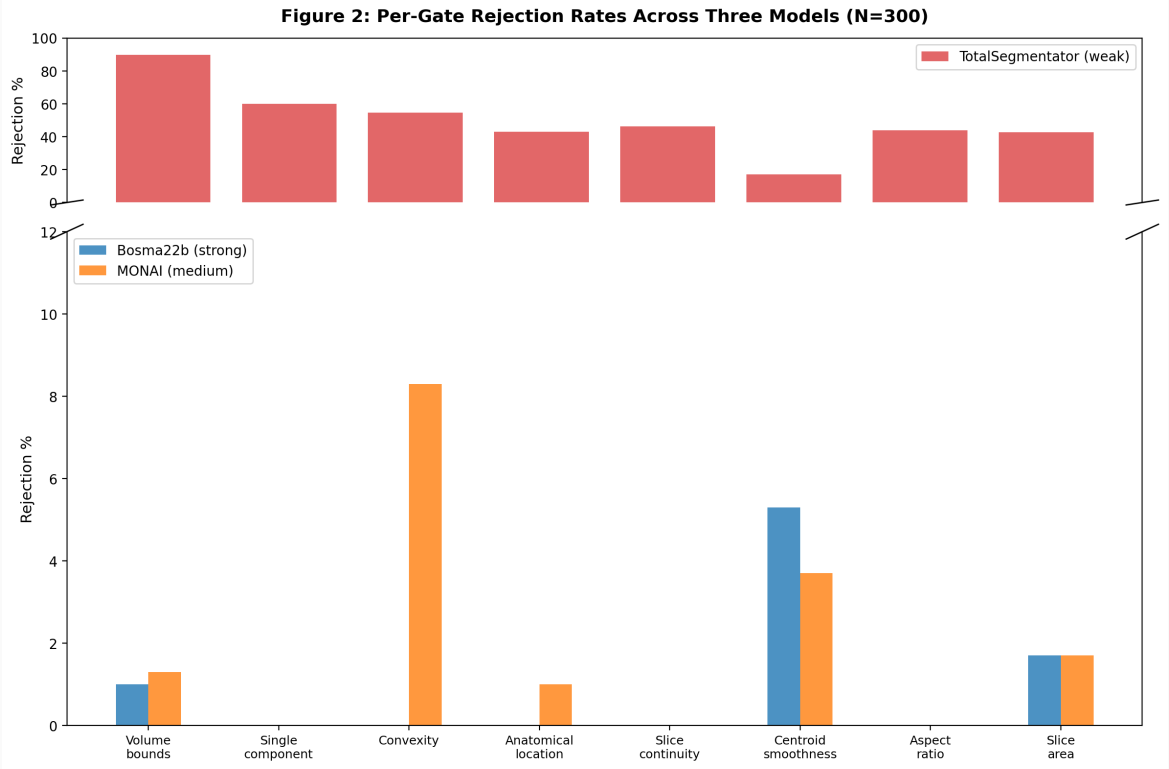


Figure 2. Per-gate rejection rates across three models (N=300 common cases). Same gates, same thresholds. Top panel: TotalSegmentator (weak). Bottom panel: Bosma22b (strong) and MONAI (medium).

4. Results

4.1 Gate Generalization (PROMISE12 → PI-CAI)

Gates calibrated on 50 PROMISE12 expert segmentations [8] generalized to 1,500 PI-CAI [10] Bosma22b masks with a 6.3% overall rejection rate (95/1,500). Four of eight gates were dormant on this strong model (single_component, anatomical_location, slice_continuity, aspect_ratio, all with zero rejections). The primary discriminator was centroid_smoothness (55 rejections, 3.7%), catching inter-slice centroid irregularities not present in expert segmentations.

The overlap ratio decreased from $\omega = 0.50$ on PROMISE12 (N=50) to $\omega = 0.24$ on PI-CAI (N=1,500), indicating greater gate complementarity on the larger, more diverse dataset.

4.2 Four-Model Scaling

Table 1. Gate performance across four models.

Model	Quality	N	Rejected	Rate	ω
Guerbet23	Strong+	1,500	72	4.8%	0.301
Bosma22b	Strong	1,500	95	6.3%	0.240
MONAI	Medium	300	33	11.0%	0.312
TotalSegmentator	Weak	300	278	92.7%	0.767

Rejection rate scaled monotonically with model weakness, now confirmed across four models including a modern multi-center nnU-Net ensemble. The strongest model tested (Guerbet23) still triggers gates on 4.8% of cases, suggesting that residual rejections persist even with high-quality, multi-center-trained segmentation.

Table 2. Per-gate rejection rates by model.

Gate	Guerbet23 (N=1500)	Bosma22b (N=1500)	MONAI (N=300)	TotalSeg (N=300)
volume_bounds	34 (2.3%)	33 (2.2%)	4 (1.3%)	269 (89.7%)
single_component	0	0	0	180 (60.0%)
convexity	2 (0.1%)	2 (0.1%)	**25 (8.3%)**	164 (54.7%)
anatomical_location	0	0	3 (1.0%)	129 (43.0%)
slice_continuity	0	0	0	139 (46.3%)
centroid_smoothness	33 (2.2%)	55 (3.7%)	11 (3.7%)	51 (17.0%)
aspect_ratio	0	0	0	132 (44.0%)
slice_area	34 (2.3%)	35 (2.3%)	5 (1.7%)	128 (42.7%)

The two strong prostate-specific models (Guerbet23 and Bosma22b) show near-identical patterns: the same 4 gates dormant, similar rates for volume_bounds and slice_area. The main difference is centroid_smoothness: Guerbet23's multi-center ensemble training produces smoother centroid trajectories (2.2% vs 3.7%), reducing the primary discriminator. The convexity gate differentiates the medium-quality model: MONAI produces less smooth boundaries (8.3% vs ≤0.1% for both strong models).

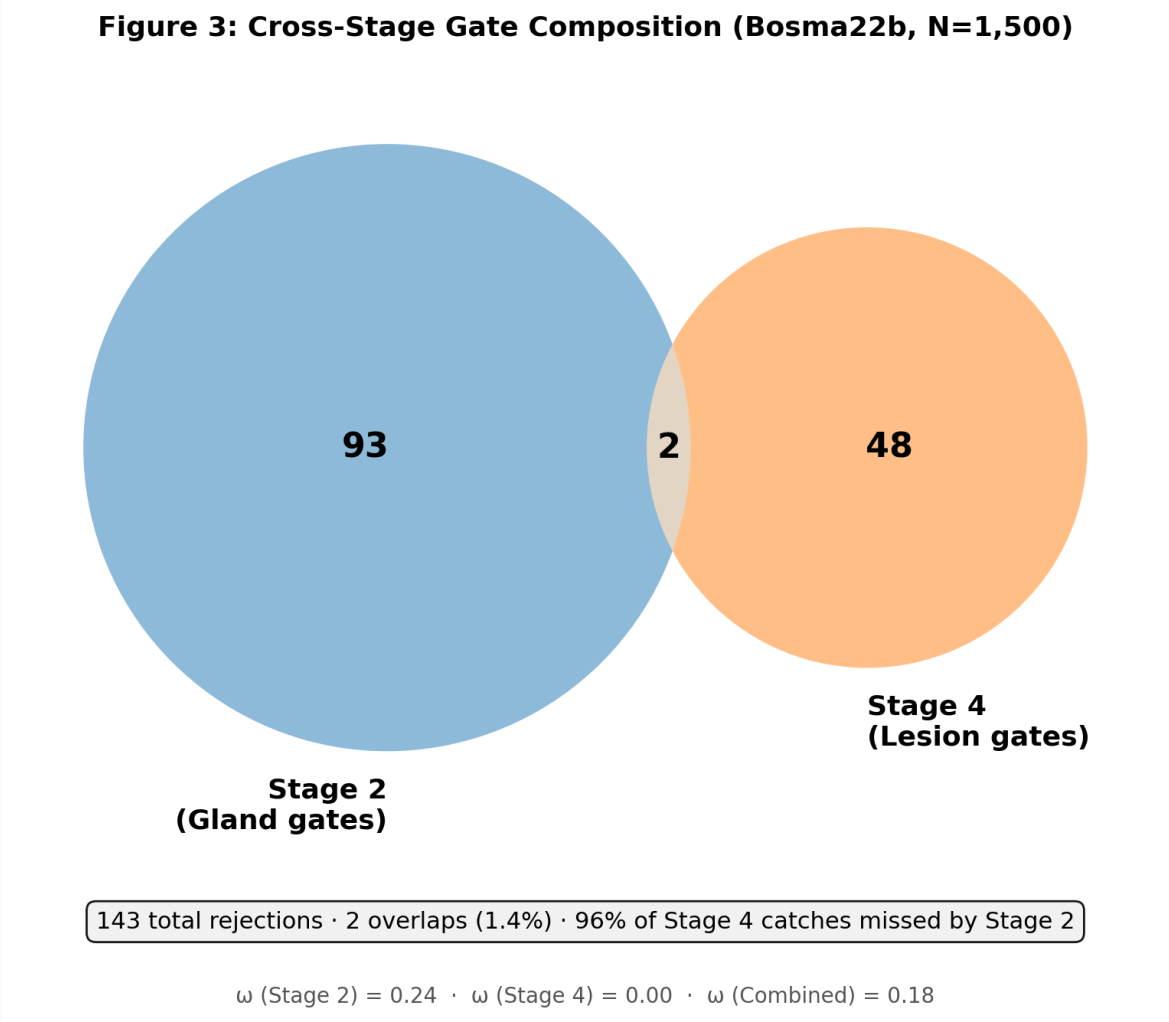


Figure 3. Cross-stage gate composition (Bosma22b, N=1,500). 93 cases caught by Stage 2 only, 48 by Stage 4 only, 2 by both. Near-zero overlap confirms cross-stage complementarity.

4.3 Cross-Stage Composition

Table 3. Stage 2 + Stage 4 composition (N = 1,500).

Metric	Bosma22b S2	Bosma22b S4	Bosma22b Combined	Guerbet23 S2	Guerbet23 S4	Guerbet23 Combined
Cases rejected	95 (6.3%)	50 (3.3%)	**143 (9.5%)**	72 (4.8%)	47 (3.1%)	**117 (7.8%)**
Unique to stage	93	48	—	70	45	—
Overlap	—	—	2	—	—	2

This is the central finding, now confirmed across two strong models. Stage 4 catches 48 cases (Bosma22b) and 45 cases (Guerbet23) not detected by any Stage 2 gate. Even with the stronger

Guerbet23 ensemble, which reduces Stage 2 rejections by 24%, Stage 4 catches nearly the same number of unique failures. The 2-case overlap in both models (1.4% and 1.7% of combined rejections) means the two stages verify nearly disjoint properties: gland shape versus lesion-gland spatial consistency.

Table 4. Overlap ratio across gate systems.

System	Bosma22b ω	Guerbet23 ω
Stage 2 alone (8 gates)	0.24	0.30
Stage 4 alone (3 gates)	0.00	0.00
Combined (11 gates)	0.18	—

Within Stage 2, the volume_bounds and slice_area gates share catches (both flag oversized prostates), producing moderate redundancy. Across stages, gates are nearly perfectly complementary in both models.

Marginal contributions in the 11-gate system (Bosma22b):

Gate	Stage	Unique catches	Total catches
centroid_smoothness	2	53	55
lesion_in_gland	4	47	49
slice_area	2	6	35
volume_bounds	2	4	33
lesion_volume	4	1	1
convexity	2	1	2

The two most valuable gates by unique catches, centroid_smoothness and lesion_in_gland, operate at different pipeline stages, together accounting for 100 of 143 total rejections (Bosma22b) and a similar proportion with Guerbet23.

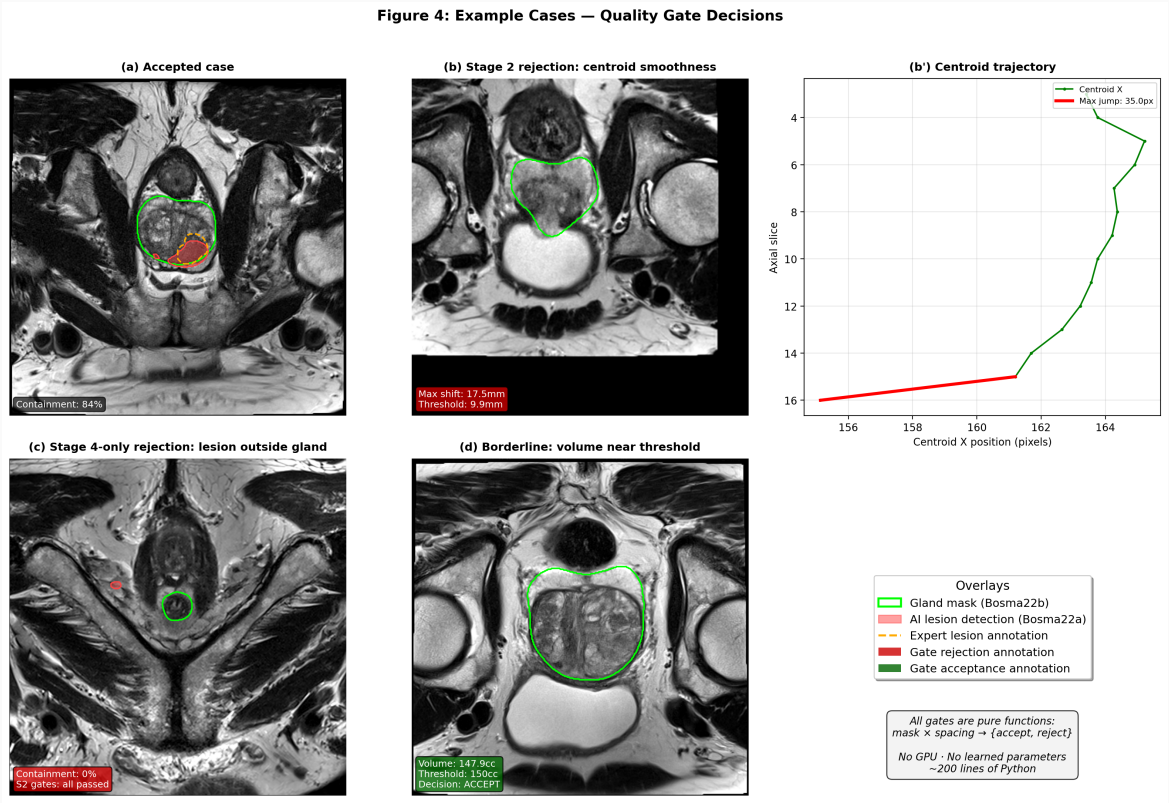


Figure 4. Example cases. (a) Accepted: plausible gland, lesion inside. (b) Stage 2 rejection: erratic centroid trajectory. (b') Centroid X position across slices showing the jump. (c) Stage 4-only rejection: gland passes all Stage 2 gates, but lesion is entirely outside the gland mask. (d) Borderline: volume 141.9cc accepted against 150cc threshold.

4.4 Verification Amplification

Table 5. Gate filtering vs random case removal: downstream containment improvement.

Filtering	N removed	Containment Δ	Random 95% CI	Exceeds?
S2 only (Bosma22b)	7	+0.006	[−0.004, +0.006]	No
S4 only (Bosma22b)	27	**+0.063**	[−0.010, +0.012]	**Yes**
Combined S2+S4	32	**+0.063**	[−0.011, +0.013]	**Yes**
MONAI (S2)	7	**+0.070**	[−0.035, +0.042]	**Yes**
TotalSegmentator (S2)	44	**+0.433**	[−0.135, +0.365]	**Yes**

For three of five configurations, gate filtering improved downstream containment significantly more than removing the same number of cases at random. The effect size scaled with model error rate: TotalSegmentator (+0.433) >> MONAI (+0.070) >> Bosma22b (+0.006).

Stage 2 gates alone did not reach significance on Bosma22b: only 7 of 220 expert-annotated cases were rejected, and the strong model produces so few gland-level failures that there is little room for selective improvement. Stage 4 gates reached significance by catching a different failure mode, lesion-gland spatial inconsistency, that is not captured by the gland-shape gate family evaluated here. This illustrates the practical value of cross-stage composition: even when single-stage QC has limited power (because the model is good enough to rarely fail on that dimension), adding QC at a different stage can catch failures along a different dimension.

We note that this result (removing structurally inconsistent cases improves metrics) is expected if the gates are working correctly. The scientific question is whether the improvement exceeds random removal, which it does for containment but not for sensitivity. This is consistent with the nature of the gates: they primarily catch spatial/structural inconsistencies (affecting containment) rather than detection sensitivity failures.

4.5 Cascading Failure Analysis

Gate-rejected cases showed degraded downstream lesion detection metrics across all three models:

Table 6. Downstream metrics in gate-rejected vs gate-accepted cases (Mann-Whitney U, one-sided).

Model	N total (N rej)	Significant metric	p	Effect size r
Bosma22b	220 (7)	Sensitivity	0.037	0.40
MONAI	47 (7)	Containment	0.005	0.62
TotalSegmentator	47 (44)	Containment	0.028	0.55

The failure mode differs by model tier. Bosma22b (strong) produces boundary failures: the gland is approximately correct but imprecise, degrading lesion sensitivity. MONAI and TotalSegmentator produce coverage failures: the gland mask misses the anatomical region, degrading lesion containment. This suggests that the same gates interact differently with different model failure modes, even though the gate code is identical.

The Bosma22b result (7 rejected cases out of 220 with lesion data) should be interpreted cautiously given the small rejection count; the $p = 0.037$ is nominally significant but statistically fragile.

4.6 Clinical Correlation

Among 1,500 PI-CAI Bosma22b cases, gate-rejected cases had lower csPCa prevalence:

Group	csPCa+	Total	Prevalence
Accepted	410	1,405	29.2%
Rejected	15	95	15.8%

Fisher's exact test: OR = 2.20, $p = 0.005$.

This association is notable but should be interpreted with care. The lower cancer prevalence in rejected cases may reflect genuine enrichment (structurally abnormal glands are less likely to harbor csPCa), but could also reflect confounds: the Bosma22b model may segment non-cancer glands less carefully, or the PI-CAI dataset may have systematic differences between cancer and non-cancer imaging. We present this as an association rather than a safety guarantee; the fact that some rejected cases (15 of 95) do contain csPCa underscores that gate rejection routes cases for review, not discard.

4.7 Failure Mode Taxonomy

Gate rejections cluster into distinct clinical failure categories, with different models exhibiting different dominant modes:

Table 7. Failure categories across models.

Failure Category	Gates	Bosma22b	Guerbet23
Structural inconsistency	single_component, slice_continuity	0	0
Boundary irregularity	convexity, centroid_smoothness	56	34
Size anomaly	volume_bounds, slice_area	39	39
Spatial misplacement	anatomical_location, aspect_ratio	0	0
Lesion-gland mismatch	lesion_in_gland, lesion_volume, lesion_count	—	47

Both strong models show zero structural inconsistency and spatial misplacement failures; these categories activate only on weaker models (MONAI: 3 spatial; TotalSegmentator: all categories active). The dominant failure modes for strong models are boundary irregularity (centroid smoothness) and size anomaly, both representing empirical plausibility violations rather than hard structural failures. Lesion-gland mismatch constitutes an entirely separate failure dimension accessible only through cross-stage gating.

4.8 Threshold Sensitivity

To test whether the main conclusions depend on specific threshold choices, we perturbed all empirical gate thresholds by ±10% and ±20% and re-evaluated Stage 2 rejection rates and cross-stage composition.

Table 8. Threshold sensitivity (Bosma22b, N = 1,500).

Perturbation	S2 rejected	S4-only catches	Cross-stage overlap
−20%	241 (16.1%)	47	1.0%
−10%	146 (9.7%)	48	1.0%
Baseline	95 (6.3%)	48	1.4%
+10%	69 (4.6%)	48	1.7%
+20%	146 (9.7%)	43	3.7%

The perturbation direction depends on gate type: for upper-bound gates (volume, slice area), +20% loosens the threshold and reduces rejections; for lower-bound gates (convexity), +20% tightens the threshold and increases rejections. Stage 2 rejection rates are sensitive to threshold perturbation, as expected since empirical bounds are by definition tunable. However, the cross-stage complementarity result is robust: Stage 4 unique catches remain between 43 and 48 across all perturbation levels, and cross-stage overlap never exceeds 3.7%. This stability arises because Stage 4 gates test a different property (lesion-gland spatial consistency) that is orthogonal to the gland-shape properties tested by Stage 2 thresholds.

The +20% perturbation produces an anomalous spike in convexity rejections (2→113) because the tightened lower bound (0.80→0.96) crosses a dense population cluster. This illustrates why threshold calibration matters for individual gate behavior, while confirming that cross-stage composition is robust to such variation.

4.9 Operational Burden and Failure Interception

From a clinical deployment perspective, the key tradeoff is review workload versus failure interception. Here, a **structural failure** means a case in which the expert lesion is poorly covered by the gland mask, defined as expert lesion containment below 80%. Among 220 expert-annotated cases, 78 (Bosma22b) and 72 (Guerbet23) meet this criterion.

Table 9. Operational burden: same-model comparison of S2-only vs S2+S4 gating.

Model	Gate Config	Cases flagged	Review rate	Structural failures intercepted	Interception rate
Bosma22b	S2-only	95	6.3%	3 of 78	3.8%
Bosma22b	S2+S4	142	9.5%	26 of 78	**33.3%**
Guerbet23	S2-only	72	4.8%	3 of 72	4.2%
Guerbet23	S2+S4	117	7.8%	25 of 72	**34.7%**

The pattern is consistent across both models: S2-only gating catches ~4% of structural failures regardless of model quality. Adding Stage 4 gates increases the review workload by ~3 percentage

points while increasing failure interception to ~34%, an 8x improvement in interception rate for a modest increase in review burden.

This improvement arises because structural failures in containment are primarily caused by lesion-gland spatial inconsistency, which the S2 gate family (testing gland shape alone) does not capture. S4 gates directly test for this failure mode, explaining both the large interception gain and the consistency across models.

Among flagged cases, rejected cases showed markedly lower mean containment (Guerbet23: 0.44 rejected vs 0.87 accepted; Bosma22b: 0.64 rejected vs 0.81 accepted), confirming that gate rejections are enriched for clinically meaningful structural inconsistencies rather than merely unusual-but-acceptable outputs.

4.10 Cross-Domain Validation: Liver and Kidney

To test whether cross-stage complementarity is specific to prostate or reflects a general property of multi-stage verification, we applied the same gate architecture to two additional domains: liver tumor segmentation (CT, N=131) and kidney tumor segmentation (CT, N=489). The same gate family was used with domain-specific threshold calibration and gate selection; no gate code was modified between domains.

Table 10. Cross-stage composition across three domains (strong model in each).

Metric	Prostate (Bosma22b)	Liver (nnU-Net)	Kidney (nnU-Net CV)
N	1,500	131	489
S1 rejected	95 (6.3%)	0 (0.0%)	46 (9.4%)
S2 rejected	50 (3.3%)	8 (6.1%)	5 (1.0%)
S2-only catches	48	8	4
Cross-stage overlap	2 (1.4%)	0 (0.0%)	1 (2.0%)

The liver result is particularly striking: nnU-Net produces zero organ-level gate failures on all 131 cases, yet Stage 2 gates still catch 8 cases (6.1%) where the liver shape is anatomically plausible but the tumor prediction is implausible. All 8 catches are S2-only; no amount of additional organ-shape gates could detect these failures. The kidney experiment uses proper cross-validation (each case predicted by the fold that did not train on it), confirming the pattern with no train-on-train contamination.

Cross-stage overlap is 0–2% in all three domains. Not all gates apply to all organs: connectivity was disabled for liver (annotation artifact) and kidney (paired organ), and aspect ratio was disabled for kidney (bilateral mask spans full abdomen width). The gate *family* is domain-agnostic but gate *selection* is domain-specific: thresholds and applicability depend on anatomy, while the gate interface and composition logic are universal.

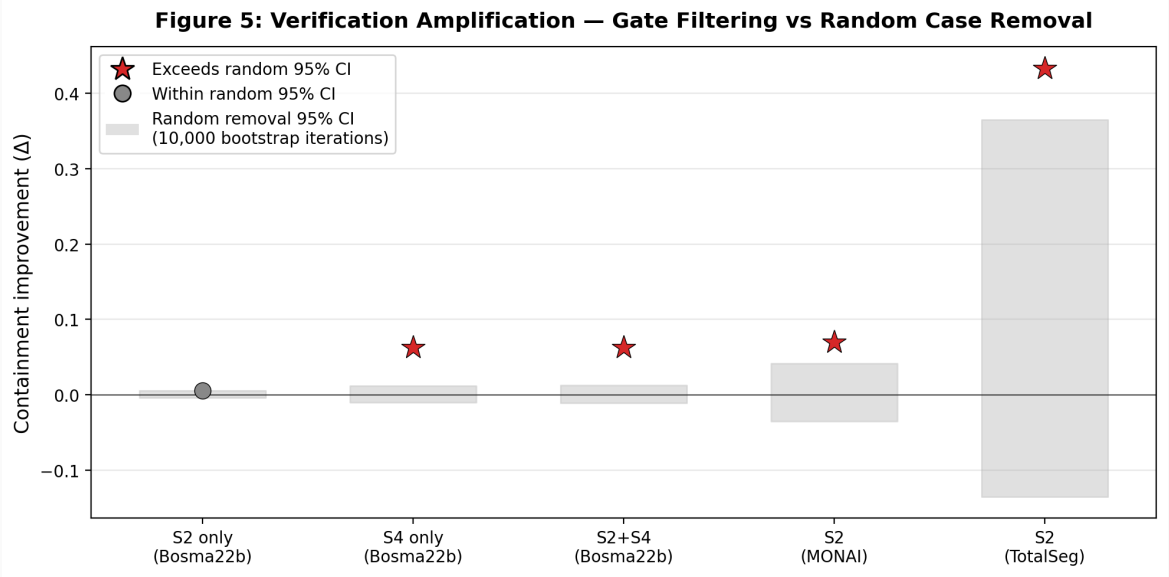


Figure 5. Verification amplification. Red stars indicate gate filtering improved containment beyond the 95% CI of random case removal (gray bands, 10,000 bootstrap iterations). Effect size scales with model weakness.

5. Discussion

5.1 What Gates Add to Existing Practice

Deterministic QC checks for AI outputs are not a new concept; they are implicit in any clinical workflow where a radiologist reviews AI-generated contours before acting on them. What this work quantifies is the **composition effect**: adding a second stage of QC catches failures not detected by any first-stage gate, with near-zero redundancy. The 48 cases caught uniquely by Stage 4 gates were not detected by any gate in the Stage 2 family evaluated here, because the gland shape itself was anatomically plausible. These failures only became visible when the lesion mask was checked against the gland mask.

This composition effect has a practical engineering implication: the marginal cost of adding gates at each pipeline stage is low (each gate is a few lines of code and requires no training), while the marginal benefit can be substantial (48 new catches from 3 additional gates, in this case).

5.2 Relationship to Model Improvement

Our results do not suggest that gates are a substitute for model improvement. A prostate-specific model (7% rejection) is clearly preferable to a multi-organ model (93% rejection), and no amount of QC gating makes a 93%-rejected pipeline clinically viable. What the results do suggest is that model improvement and QC gating address different dimensions of reliability:

- **Model improvement** reduces the frequency of failures (fewer anatomically implausible outputs).

- **QC gating** reduces the impact of residual failures (catching implausible outputs before they propagate).

In this pipeline, even the strongest model (Guerbet23, a multi-center nnU-Net ensemble) produces gland segmentations that fail basic plausibility checks on 4.8% of cases, and lesion-gland inconsistencies on an additional 3.1% of cases. These are cases where post-hoc QC provides a different kind of value than model improvement alone, catching failures after they occur rather than preventing them.

5.3 Complementarity with Uncertainty Estimation

Deterministic gates and stochastic uncertainty estimation address different failure modes:

	Deterministic Gates	**Uncertainty Estimation**
Catches	Structural violations, spatial inconsistency	Boundary imprecision, distributional shift
Requires	Anatomical knowledge, threshold calibration	Training data, Monte Carlo sampling
Cost	Negligible (pure functions)	Significant (multiple forward passes)
Model dependence	None (tests output only)	Requires model internals or outputs
Guarantee	Deterministic behavior; explicit coverage of specified failure classes; clinical correctness of individual rejections not globally guaranteed	Uncertainty is empirically calibrated; coverage depends on training distribution

A clinical verification system would ideally compose both: deterministic gates providing an interpretable QC floor, with uncertainty estimation providing additional coverage for subtler failures.

5.4 Limitations

Cross-domain validation is preliminary. The cross-stage complementarity result replicates across prostate (MRI, N=1,500), liver (CT, N=131), and kidney (CT, N=489), with near-zero cross-stage overlap in all three domains (Section 4.10). However, the liver experiment has a train-on-train limitation (pretrained model evaluated on training data), and absolute S2-only catch counts are small in liver (8) and kidney (4). Specific thresholds and gate applicability differ by anatomy, and not all gates apply to all organs.

Threshold sensitivity. While the cross-stage composition result is robust to ±20% threshold perturbation (Section 4.8), individual gate rejection rates are sensitive to threshold choices. Robustness to substantially different imaging protocols (3T vs 1.5T, endorectal coil) is not established.

Small rejection counts. Bosma22b produces so few gland-level failures that cascading failure and amplification analyses are limited by small N in the rejected group (7 cases with lesion data). The p = 0.037 sensitivity result is nominally significant but underpowered.

No expert adjudication of gate decisions. This study validates gates against automated downstream metrics (containment, sensitivity) and statistical tests, but does not include an expert review study in which radiologists evaluate whether gate-flagged cases are clinically judged to be unacceptable. The gates detect structural inconsistency as defined by the gate criteria; whether those criteria align with clinical judgment on every case remains unvalidated. An expert adjudication study (in which radiologists review a sample of accepted and rejected cases and rate clinical acceptability) would be the most direct test of gate utility and is the natural next step for this work.

Expert annotation coverage. Expert lesion delineations are available for only 220 of 1,500 cases, limiting the scope of downstream impact analysis.

Two-stage composition. Stages 1, 3, and 5 remain ungated. Whether adding gates at additional stages continues to provide complementary coverage is plausible but undemonstrated.

Binary decisions. Gates produce accept/reject without graded quality scores. Cases near thresholds may be unnecessarily flagged. Soft-threshold extensions could provide continuous quality estimates while preserving the binary decision for operational routing.

Empirical thresholds are not universal guarantees. The empirical plausibility bounds (volume, convexity, centroid smoothness, etc.) are derived from observed clinical ranges and may not hold for all populations. Rare but genuine anatomical variants could trigger false rejections. Hard structural constraints (connectivity, slice continuity) are more robust to population variation but are also less discriminating on strong models.

---

6. Conclusion

We demonstrate that 11 deterministic quality gates (approximately 200 lines of Python, requiring no GPU or learned parameters) provide interpretable, auditable failure detection across two stages of a prostate cancer AI pipeline. The most notable finding is the cross-stage composition effect: lesion-stage gates catch 48 cases (Bosma22b) and 45 cases (Guerbet23) missed by all gland-stage gates, with near-zero overlap. This result holds even for a modern multi-center nnU-Net ensemble with the lowest gland-level rejection rate (4.8%). Gate filtering selectively improves downstream metrics beyond random case removal for models with moderate to high error rates.

These gates are not a replacement for model improvement or uncertainty estimation. They are a complementary, low-cost layer of QC that catches gross structural failures before they propagate through multi-stage pipelines. For clinical AI deployment, this kind of interpretable, deterministic failure detection may be an important companion to model development, especially in pipelines where reliability depends not only on model quality but on how the overall process is decomposed and verified.

Figures

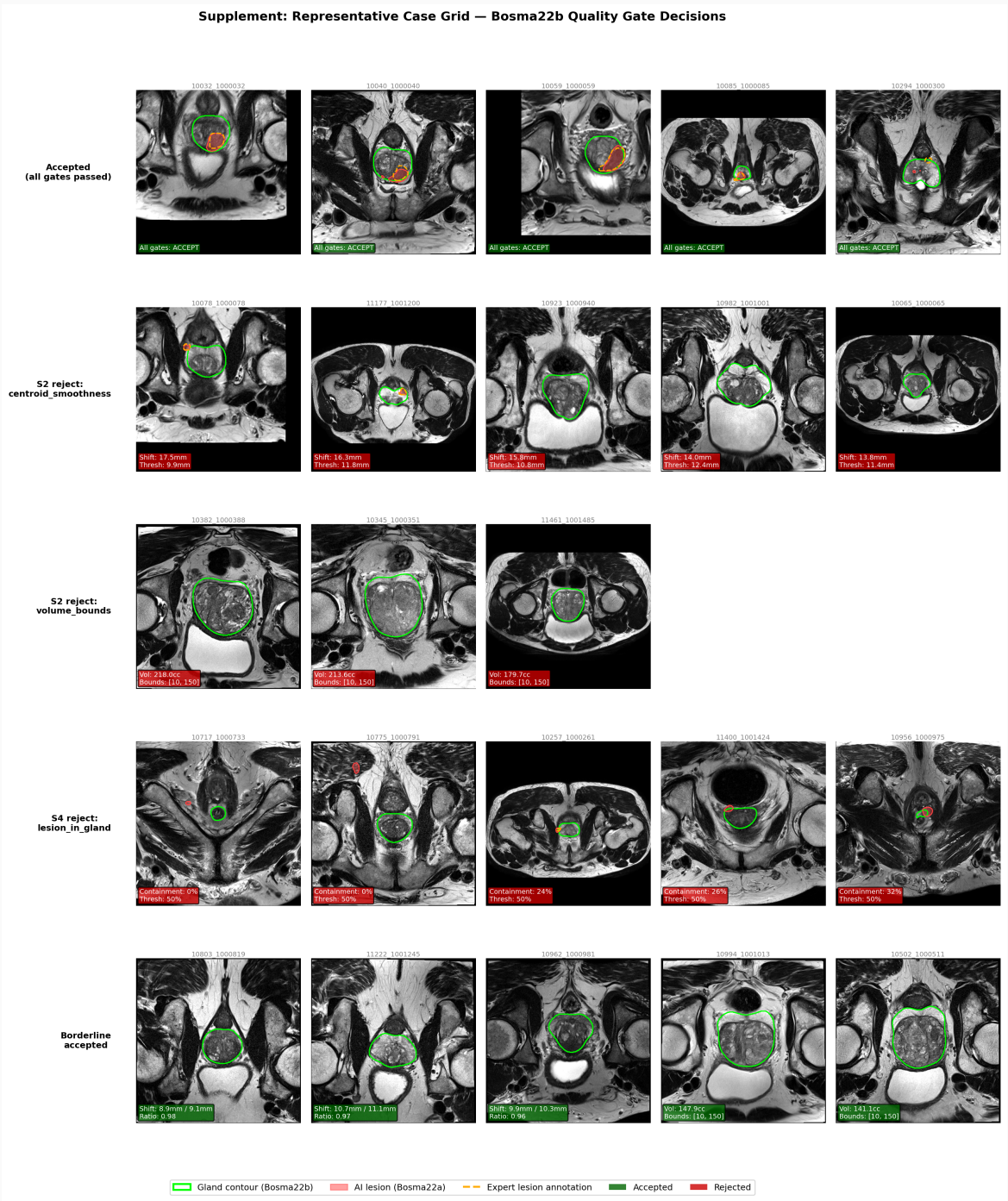
Figure 1. Pipeline architecture showing five stages with quality gates at Stage 2 and Stage 4.

Figure 2. Three-model gate scaling. Per-gate rejection rates for Bosma22b, MONAI, and TotalSegmentator.

Figure 3. Cross-stage composition Venn diagram: Stage 2 rejections (93 unique), Stage 4 rejections (48 unique), overlap (2 cases).

Figure 4. Example cases. (a) Accepted case. (b) Stage 2 rejection: centroid smoothness failure. (c) Stage 4 rejection missed by Stage 2: plausible gland shape, lesion outside boundary. (d) Borderline case near volume threshold.

Figure 5. Verification amplification. Gate filtering vs 10,000 random-removal iterations per model.



Supplementary Figure. Case grid showing representative examples across gate decision categories. Each panel shows a T2W axial slice with gland mask contour (green) and lesion overlay (red, where present). Annotations indicate gate values and thresholds. Top row: accepted cases. Rows 2-4: gate rejections by type. Bottom row: borderline cases near thresholds.

References

1. Y. Gal and Z. Ghahramani, “Dropout as a Bayesian Approximation: Representing Model Uncertainty in Deep Learning,” in *Proceedings of the 33rd International Conference on Machine Learning (ICML)*, PMLR 48, pp. 1050–1059, 2016.
2. B. Lakshminarayanan, A. Pritzel, and C. Blundell, “Simple and Scalable Predictive Uncertainty Estimation using Deep Ensembles,” in *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 30, pp. 6402–6413, 2017.
3. G. Wang, W. Li, M. Aertsen, J. Deprest, S. Ourselin, and T. Vercauteren, “Aleatoric uncertainty estimation with test-time augmentation for medical image segmentation with convolutional neural networks,” *Neurocomputing*, vol. 338, pp. 34–45, 2019.
4. R. Robinson, O. Oktay, W. Bai, V. V. Valindria, M. M. Sanghvi, N. Aung, J. M. Paiva, F. Zemrak, K. Fung, E. Lukaschuk, S. E. Petersen, D. Rueckert, and B. Glocker, “Real-time prediction of segmentation quality,” in *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, LNCS 11073, pp. 578–585, 2018.
5. T. Kohlberger, V. Singh, C. Alvino, C. Bahlmann, and L. Grady, “Evaluating segmentation error without ground truth,” in *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, LNCS 7510, pp. 528–536, 2012.
6. V. V. Valindria, I. Lavdas, W. Bai, K. Kamnitsas, E. O. Aboagye, A. G. Rockall, D. Rueckert, and B. Glocker, “Reverse classification accuracy: predicting segmentation performance in the absence of ground truth,” *IEEE Transactions on Medical Imaging*, vol. 36, no. 8, pp. 1597–1606, 2017.
7. J. S. Bosma, A. Saha, M. Hosseinzadeh, I. Slootweg, M. de Rooij, and H. Huisman, “Semisupervised Learning with Report-guided Pseudo Labels for Deep Learning–based Prostate Cancer Detection Using Biparametric MRI,” *Radiology: Artificial Intelligence*, vol. 5, no. 5, e230031, 2023.
8. G. Litjens, R. Toth, W. van de Ven, C. Hoeks, S. Kerkstra, B. van Ginneken, G. Vincent, G. Guillard, N. Birbeck, J. Zhang, R. Strand, F. Malmberg, Y. Ou, C. Davatzikos, M. Kirschner, F. Jung, J. Yuan, W. Qiu, Q. Gao, P. E. Edwards, B. Maan, F. van der Heijden, S. Ghose, J. Mitra, J. Dowling, D. Barratt, H. Huisman, and A. Madabhushi, “Evaluation of prostate segmentation algorithms for MRI: the PROMISE12 challenge,” *Medical Image Analysis*, vol. 18, no. 2, pp. 359–373, 2014.
9. J. Wasserthal, H.-C. Breit, M. T. Meyer, M. Pradella, D. Hinck, A. W. Sauter, T. Heye, D. T. Boll, J. Cyriac, S. Yang, M. Bach, and M. Segeroth, “TotalSegmentator: Robust Segmentation of 104 Anatomic Structures in CT Images,” *Radiology: Artificial Intelligence*, vol. 5, no. 5, e230024, 2023.
10. A. Saha, J. S. Bosma, J. J. Twilt, B. van Ginneken, A. Bjartell, A. Padhani, D. Bonekamp, G. Villeirs, G. Salomon, G. Giannarini, J. Fütterer, H. Huisman, and M. de Rooij, “Artificial Intelligence and Radiologists in Prostate Cancer Detection on MRI (PI-CAI): An International,

Paired, Non-inferiority, Confirmatory Study," *The Lancet Oncology*, vol. 25, no. 7, pp. 879–887, 2024.

11. F. Isensee, P. F. Jaeger, S. A. A. Kohl, J. Petersen, and K. H. Maier-Hein, "nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation," *Nature Methods*, vol. 18, no. 2, pp. 203–211, 2021.
12. Guerbet, "Prostate Segmentation," Grand Challenge algorithm, 2023. nnU-Net 5-fold ensemble trained on PI-CAI (1,500 bpMRI scans, 11 centers, 7 scanners). Available: https://grand-challenge.org/algorithms/guerbet-prostate-segmentation/ (accessed March 2026).