SHIFTSAFE-QA: SOURCE-SUPPORTED CONSERVATIVE QUALITY ASSESSMENT UNDER ACQUISITION SHIFT

Linkun Yu[1,*], *Xiao Zhang*[2], *Jiaqi Zhang*[3], *Ping Guo*[4], *Zehua Yang*[4], *Fred Sun*[4]

[1]Graduate School of Fundamental Science and Engineering, Waseda University, Tokyo, Japan
[2]Graduate School of Informatics, Nagoya University, Nagoya, Japan
[3]School of Computer Science and Telecommunication Engineering, Jiangsu University, China
[4]Graduate School of Information, Production and Systems, Waseda University, Fukuoka, Japan

ABSTRACT

Case-level segmentation quality assessment (SQA) can support automated acceptance, but a score calibrated on one acquisition protocol can become overconfident after acquisition shift. We introduce ShiftSafe-QA, an inference-time correction for a frozen Dice predictor. It subtracts a source-supported penalty from the predicted Dice, where support is measured in the quality-feature space of segmentation probabilities, masks, and image-mask appearance. The only penalty weight is selected using held-out source acquisitions with synthetic intensity and contrast changes; no target labels or target tuning are used. In prostate MRI, the ordinary image-mask predictor accepts 77 of 150 PROMISE12 cases at a 0.80 score threshold, including 30 failed masks. ShiftSafe-QA accepts 26 cases with no failures, while a hard OOD gate accepts only three. On more severe paired ADC and Prostate158 shifts, ShiftSafe-QA abstains rather than retaining unsafe source-threshold acceptances. These results position source support as a practical signal for conservative SQA under acquisition shift.

***Index Terms*—** Segmentation quality assessment, medical image segmentation, uncertainty, acquisition shift, selective prediction.

1. INTRODUCTION

Medical segmentation systems are increasingly evaluated not only by their average overlap, but also by whether an individual predicted mask is trustworthy enough to retain without review. Segmentation quality assessment (SQA) addresses this case-level question by predicting an unavailable Dice score from a model output and its input image [1, 2, 3, 4, 5]. Such a predicted score can rank masks, triage difficult cases, or define an automated acceptance rule.

The key operational difficulty is that the acceptance score is normally learned and calibrated on a source acquisition. Scanner, sequence, reconstruction, and contrast changes can alter the image and probability-map statistics without changing the downstream decision threshold. Domain shifts of this type are well documented in medical segmentation [6, 7]; however, a high quality score alone gives no indication that the new case resembles the source conditions under which that score was learned. Standard uncertainty summaries can be similarly overconfident outside their training distribution [8, 9, 10, 11].

We ask a narrower question: *can a frozen quality score be made conservative when its feature representation is weakly supported by source data?* We answer this with ShiftSafe-QA, a small post-hoc method that retains an existing SQA predictor and discounts its predicted Dice according to a robust source-support distance. The design deliberately does not adapt the segmentation network or use target labels. Its single penalty is fixed using held-out source acquisitions with intensity/contrast perturbations, thereby separating calibration from all external targets.

Our contributions are threefold. First, we formulate source-supported SQA as a continuous conservative correction rather than a binary OOD rejection rule. Second, we provide a source-only selection protocol for its correction strength. Third, across controlled T2-to-ADC shift and two external prostate MRI cohorts, we show that the correction removes unsafe PROMISE12 acceptances at materially higher coverage than a hard source-support gate, while correctly abstaining under more severe shifts.

2. SHIFTSAFE-QA

Figure 1 provides an overview of the source-only calibration and target-time safeguard.

2.1. Frozen quality prediction and source support

Let x be an image, $p(x)$ a frozen segmentation probability map, and $m(x)$ its predicted mask. An ordinary SQA model maps a case feature vector $z(x, p, m)$ to a Dice estimate $\hat{q} \in [0, 1]$. We use a ridge-regularized linear image-mask regres-

* Corresponding author.

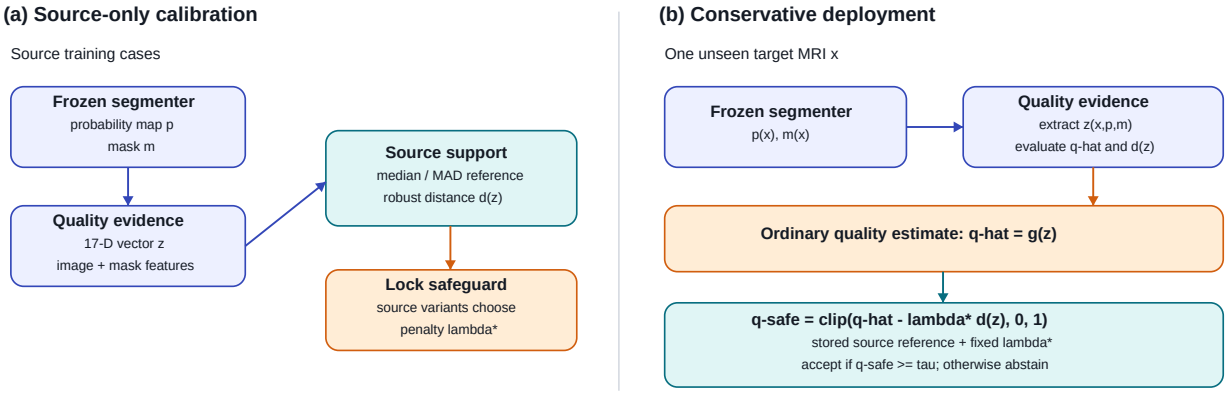

Fig. 1. ShiftSafe-QA overview. **(a)** Source-only calibration fits the frozen quality predictor, stores robust feature reference statistics, and locks the penalty using acquisition-like variants from a disjoint source split. **(b)** At deployment, the same frozen segmentation and quality evidence yield an ordinary Dice estimate and source-support distance; the conservative score discounts unsupported evidence before the fixed accept-or-abstain decision.

sor, trained only on source cases, so the quality model and the segmentation network remain frozen at deployment. This deliberately lightweight choice isolates the proposed correction from segmentation architecture and avoids retraining the segmenter.

For the image-mask predictor, z contains 17 case-level statistics. Ten quality features summarize mean and boundary entropy, confidence and foreground-probability moments, predicted-mask area, boundary density, compactness, connected components, and test-time-augmentation consistency. Seven appearance features summarize image intensity globally, inside the mask, in the local background, and on the predicted boundary, including mask-background contrast. Every feature is computed after a mask is produced, so the correction introduces no additional image encoder or segmentation pass.

The construction ties source support to the evidence actually consumed by the Dice predictor, rather than to a global image-level domain label. A case can therefore look visually familiar yet have an atypical mask–probability–appearance configuration; conversely, a global acquisition change is not penalized unless it changes this quality-relevant evidence.

Let $\mathcal{S} = \{z_i\}_{i=1}^{n}$ be source-training quality features. For each coordinate j, we compute a robust location $\tilde{z}_j = \mathrm{median}_{z\in\mathcal{S}} z_j$ and scale $s_j = 1.4826\,\mathrm{MAD}_{z\in\mathcal{S}}(z_j) + \epsilon$. The source-support distance of a test case is

$$d(z) = \frac{1}{D}\sum_{j=1}^{D}\left|\frac{z_j - \tilde{z}_j}{s_j}\right|. \tag{1}$$

Unlike a detector trained to classify a particular target domain, d only measures whether the feature evidence for a quality estimate is represented by source cases. Median/MAD scaling makes the calculation stable for heterogeneous feature units and avoids fitting a full covariance matrix from a small source cohort.

2.2. Conservative score and source-only selection

ShiftSafe-QA reports the conservative estimate

$$q_{\mathrm{safe}} = \mathrm{clip}(\hat{q} - \lambda d(z), 0, 1), \tag{2}$$

where $\lambda \geq 0$ trades ordinary score against unsupported evidence. A case is accepted only if $q_{\mathrm{safe}} \geq \tau$, with $\tau = 0.80$. We define an accepted failure as an accepted case whose true Dice is below 0.75. This separates a stringent acceptance score from the failure label used to audit the decision.

The penalty is selected without target labels, target batches, or a target validation split. We reserve a patient-disjoint source calibration split $\mathcal{A}_{\mathrm{cal}}$, create fixed intensity and contrast variants for every calibration image, and choose

$$\lambda^* = \arg\max_{\lambda}\ \mathrm{Coverage}(\lambda; \mathcal{A}_{\mathrm{cal}}) \quad \text{s.t.} \quad \mathrm{AFR}(\lambda; \mathcal{A}_{\mathrm{cal}}) \leq 0.05, \tag{3}$$

where AFR is accepted-failure rate. This is a source-only robustness requirement: it selects the most permissive correction that remains conservative under acquisition-like perturbations; ties favor the smaller penalty, and the selected value is locked before every external target is scored. As a matched control, a global offset $\mathrm{clip}(\hat{q} - c, 0, 1)$ selects c by the same rule but ignores support distance. A hard OOD gate rejects samples above the 95th percentile of a source Mahalanobis distance, following conventional source-support screening [8, 12].

The correction has a deliberately asymmetric role. When the source support is high, $d(z)$ is small and the reported score remains close to the ordinary Dice estimate. When the same regressor encounters unsupported evidence, the score

is discounted before applying the fixed acceptance threshold. ShiftSafe-QA therefore differs from recalibration on a new domain: it makes no claim that the adjusted value is an unbiased target-domain Dice estimate. Its intended use is narrower—to avoid treating a source-calibrated score as an automatic approval when the evidence underlying that score has moved away from source support.

2.3. Source-Only Selection and Deployment Protocol

The complete procedure separates an offline source-only stage from an online case-level stage. Offline, we train the frozen ridge regressor on source cases, store the coordinate-wise median and MAD of its 17-dimensional evidence vectors, generate calibration acquisitions from a disjoint source subset, and select λ^* once using Eq. (3). The segmentation network, quality regressor, robust statistics, penalty, and acceptance threshold are then all locked. Hence neither the identity, prevalence, nor labels of a target cohort can change the deployed rule.

Online, each case requires one frozen segmentation prediction and the same post-hoc feature extraction already required by the ordinary image-mask SQA baseline. We evaluate $\hat{q}$, compute $d(z)$ using the stored source statistics, form q_{safe}, and return either an accepted automated result or an abstention for downstream review. The correction is scalar and has no trainable target-time component, no target-batch normalization, and no iterative adaptation. This makes the decision auditable: an abstention can be attributed to either a low ordinary quality estimate, weak source support, or their combination. We report both score accuracy and the resulting acceptance behavior because a favorable ranking metric alone does not establish that a fixed source threshold is safe after an acquisition shift.

3. EXPERIMENTS

3.1. Data and protocol

We use the Medical Segmentation Decathlon Task05 prostate T2 MRI data as source [13]. A source training split defines the segmentation and quality-feature reference; eight held-out source patients form $\mathcal{A}_{\mathrm{cal}}$, and an independent eight-patient source test split evaluates in-domain behavior. Target evaluation includes PROMISE12 whole-prostate T2 MRI [14], a paired T2-to-ADC acquisition target, and the multi-institution Prostate158 cohort [15]. The paired target holds anatomy fixed while changing contrast, isolating acquisition sensitivity from patient or annotation changes.

All methods use identical frozen segmentation predictions and the same image-mask quality regressor. We repeat the source splits with three fixed seeds and pool their cases for decision statistics. We report coverage, accepted-failure rate, Dice-prediction MAE, and risk–coverage curves. Selection of λ is performed separately in each seed using only $\mathcal{A}_{\mathrm{cal}}$ and its prescribed augmentations. Neither target labels nor target feature distributions enter model fitting, penalty selection, or threshold selection.

Table 1. Fixed-threshold acceptance at $\tau = 0.80$, pooled over three fixed splits. Entries are accepted cases / accepted failures.

Target	Ordinary QA	ShiftSafe-QA	Hard OOD gate
Source test	24 / 4	14 / 0	0 / 0
PROMISE12	77 / 30	26 / 0	3 / 0
Paired ADC	10 / 10	0 / 0	0 / 0
Prostate158	46 / 46	0 / 0	0 / 0

Table 2. PROMISE12 transfer under source-selected thresholds. Entries are accepted cases / accepted failures.

Method	Accepted / failures
Mean entropy	53 / 39
Mean confidence	52 / 38
TTA consistency	75 / 48
Structure-only	18 / 3
Global offset	0 / 0
ShiftSafe-QA	26 / 0
Hard OOD gate	3 / 0

The three targets probe complementary transfer conditions. PROMISE12 tests ordinary external-cohort transfer within T2 MRI; the paired T2-to-ADC set isolates contrast change while holding patient anatomy fixed; Prostate158 combines external acquisition and multi-institutional variation. We keep the acceptance threshold fixed across all settings. Consequently, lower coverage on a target is not hidden by retuning the decision rule: it is the operational cost of refusing to certify a prediction using a source-only score.

Table 1 summarizes the primary fixed-threshold deployment comparison. It contrasts ordinary QA, the proposed source-supported correction, and a conservative hard gate before the subsequent analyses explain the individual failure modes and penalty choices.

3.2. Baseline QA under external shift

Before applying the conservative correction, we quantify transfer of the underlying image-mask regressor on the same frozen segmentation outputs. Source-test MAE/ECE are $0.037/0.021$ over the fixed splits, whereas both absolute error and calibration worsen on all targets. PROMISE12 retains strong rank correlation, but its score scale shifts; paired ADC loses both calibration and ranking; and Prostate158 has the largest absolute error. Table 2 compares source-selected single-score rules on PROMISE12. Entropy, confidence, and TTA consistency remain unsafe; structure-only scoring is the strongest simple rule but still accepts three failures. The matched global offset and hard gate avoid failures only by accepting 0 and 3 cases, respectively; ShiftSafe-QA accepts 26.

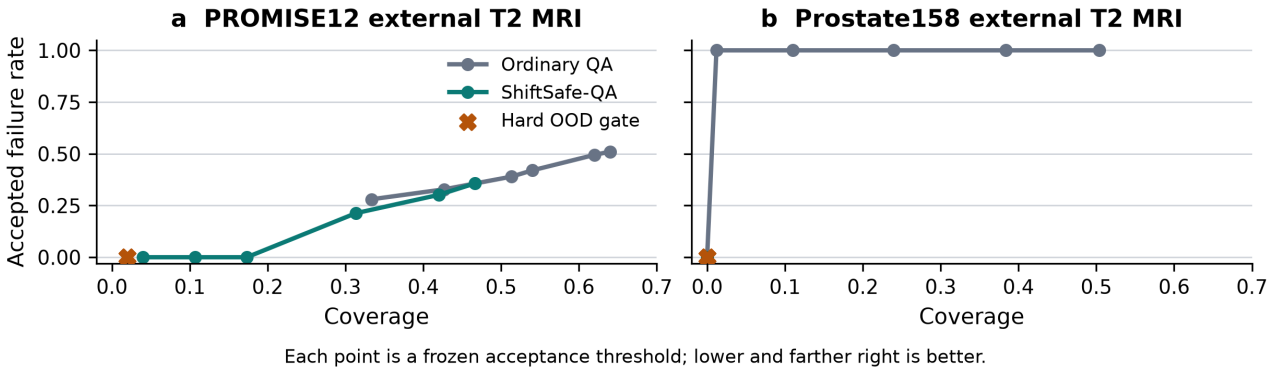

Fig. 2. Risk–coverage behavior of ordinary image-mask SQA, ShiftSafe-QA, and a hard source-support gate. **Left:** PROMISE12. Continuous source-supported correction retains appreciably more coverage than hard gating while reducing accepted-failure risk. **Right:** Prostate158. The conservative score abstains at the fixed 0.80 threshold under this severe shift rather than preserving unsafe source-threshold acceptances.

Table 3. PROMISE12 support-distance ablation at $\tau = 0.80$, pooled over three splits. Entries are accepted / failed accepted cases.

Distance	Accepted / failures
Robust L1 (default)	26 / 0
Standardized L2	27 / 0
Mahalanobis	9 / 0

3.3. Source-only penalty selection

Each penalty is selected from 40 held-out source augmentation cases, separately for every split; the selected robust-L1 penalties are 0.045, 0.080, and 0.075. The correction removes accepted failures in all three repetitions without a target-dependent choice of λ. Table 3 tests whether this result depends on the support metric. Standardized L2 is nearly tied with robust L1, whereas continuous Mahalanobis distance is much more conservative with this small source cohort.

3.4. Results

The fixed-threshold comparison exposes the deployment failure hidden by an ordinary source-calibrated score. On PROMISE12, ordinary QA accepts 77/150 cases, but 30 of these masks fail. ShiftSafe-QA accepts 26/150 with no failures. A 5,000-replicate seed-and-case bootstrap gives ordinary accepted-failure risk 0.390 (95% CI $[0.250, 0.536]$) and a ShiftSafe-minus-ordinary risk difference -0.390 (95% CI $[-0.536, -0.250]$). The hard gate also avoids failures, but retains only three cases, demonstrating why a continuous correction is preferable to rejecting all unsupported examples.

Figure 2 gives the full threshold behavior. On PROMISE12, ShiftSafe-QA lowers risk across the operating range. At $\tau = 0.80$, it reaches zero observed accepted failures.

The paired ADC and Prostate158 results identify the boundary of that decision. Ordinary QA retains accepted masks even though every accepted mask in these pooled severe-shift settings fails. On Prostate158, mean entropy, mean confidence, and TTA consistency accept 40, 43, and 30 cases, respectively, and every accepted mask fails; the simple structure-only rule abstains. ShiftSafe-QA also accepts no cases at the fixed threshold. This abstention is informative: a method that cannot establish sufficient source support should not convert its score into an automated acceptance. In particular, our method does not restore useful coverage for arbitrary external shifts; it avoids presenting unsupported quality estimates as safe decisions.

4. DISCUSSION AND CONCLUSION

ShiftSafe-QA is a post-hoc safeguard for conservative selective deployment. It does not adapt the segmenter or recalibrate Dice on an unseen target; instead, it discounts a learned quality score when the image–mask evidence lacks source support. Unlike a global score offset, the penalty is case dependent: familiar evidence remains close to the ordinary estimate, whereas unsupported evidence is made less likely to pass the fixed acceptance threshold.

The experiments separate useful ranking from a deployable acceptance rule. A score can retain ordering on an external cohort while still accepting unsafe masks at its source-calibrated threshold. ShiftSafe-QA reduces this decision error on PROMISE12 while retaining more cases than hard support gating, and abstains under the paired ADC and Prostate158 shifts when support is inadequate. The study is limited to prostate MRI, one feature family, and small source calibration splits; future work should test learned support metrics, multi-class masks, and prospective workflows in which abstention triggers review.

5. REFERENCES

[1] Alain Jungo and Mauricio Reyes, "Assessing reliability and challenges of uncertainty estimations for medical image segmentation," in *Medical Image Computing and Computer Assisted Intervention – MICCAI 2019*. 2019, pp. 48–56, Springer.

[2] Alireza Mehrtash, William M. Wells, Clare M. Tempany, Purang Abolmaesumi, and Tina Kapur, "Confidence calibration and predictive uncertainty estimation for deep medical image segmentation," *IEEE Transactions on Medical Imaging*, vol. 39, no. 12, pp. 3868–3878, 2020.

[3] Robert Robinson, Ozan Oktay, Wenjia Bai, Vanya V. Valindria, Mihir M. Sanghvi, Nay Aung, José Miguel Paiva, Filip Zemrak, Kenneth Fung, Elena Lukaschuk, Aaron M. Lee, Valentina Carapella, Young Jin Kim, Bernhard Kainz, Stefan K. Piechnik, Stefan Neubauer, Steffen E. Petersen, Chris Page, Daniel Rueckert, and Ben Glocker, "Real-time prediction of segmentation quality," in *Medical Image Computing and Computer Assisted Intervention – MICCAI 2018*. 2018, pp. 578–585, Springer.

[4] Robert Robinson, Vanya V. Valindria, Wenjia Bai, Ozan Oktay, Bernhard Kainz, Hideaki Suzuki, Mihir M. Sanghvi, Nay Aung, José Miguel Paiva, Filip Zemrak, Kenneth Fung, Elena Lukaschuk, Aaron M. Lee, Valentina Carapella, Young Jin Kim, Stefan K. Piechnik, Stefan Neubauer, Steffen E. Petersen, Chris Page, Paul M. Matthews, Daniel Rueckert, and Ben Glocker, "Automated quality control in image segmentation: Application to the UK biobank cardiovascular magnetic resonance imaging study," *Journal of Cardiovascular Magnetic Resonance*, vol. 21, pp. 18, 2019.

[5] Joris Fournel, Axel Bartoli, David Bendahan, Maxime Guye, Monique Bernard, Elisa Rauseo, Mohammed Y. Khanji, Steffen E. Petersen, Alexis Jacquier, and Badih Ghattas, "Medical image segmentation automatic quality control: A multi-dimensional approach," *Medical Image Analysis*, vol. 74, pp. 102213, 2021.

[6] Hao Guan and Mingxia Liu, "Domain adaptation for medical image analysis: A survey," *IEEE Transactions on Biomedical Engineering*, 2022.

[7] Kaiyang Zhou, Ziwei Liu, Yu Qiao, Tao Xiang, and Chen Change Loy, "Domain generalization: A survey," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, no. 4, pp. 4396–4415, 2023.

[8] Yaniv Ovadia, Emily Fertig, Jie Ren, Zachary Nado, D. Sculley, Sebastian Nowozin, Joshua V. Dillon, Balaji Lakshminarayanan, and Jasper Snoek, "Can you trust your model's uncertainty? evaluating predictive uncertainty under dataset shift," in *Advances in Neural Information Processing Systems*, 2019, vol. 32.

[9] Alex Kendall and Yarin Gal, "What uncertainties do we need in bayesian deep learning for computer vision?," in *Advances in Neural Information Processing Systems*, 2017, vol. 30.

[10] Yarin Gal and Zoubin Ghahramani, "Dropout as a bayesian approximation: Representing model uncertainty in deep learning," in *Proceedings of the 33rd International Conference on Machine Learning*, 2016, vol. 48 of *Proceedings of Machine Learning Research*, pp. 1050–1059.

[11] Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell, "Simple and scalable predictive uncertainty estimation using deep ensembles," in *Advances in Neural Information Processing Systems*, 2017, vol. 30.

[12] Dan Hendrycks and Kevin Gimpel, "A baseline for detecting misclassified and out-of-distribution examples in neural networks," in *International Conference on Learning Representations*, 2017.

[13] Michela Antonelli, Annika Reinke, Spyridon Bakas, et al., "The medical segmentation decathlon," *Nature Communications*, vol. 13, pp. 4128, 2022.

[14] Geert Litjens, Robert Toth, Wendy van de Ven, et al., "Evaluation of prostate segmentation algorithms for MRI: The PROMISE12 challenge," *Medical Image Analysis*, vol. 18, no. 2, pp. 359–373, 2014.

[15] Lisa C. Adams, Marcus R. Makowski, Günther Engel, et al., "Prostate158: An expert-annotated 3T MRI dataset and algorithm for prostate cancer detection," *Computers in Biology and Medicine*, vol. 148, pp. 105817, 2022.