Preserving Recall Calibration through Connected-Component Filtering in Polyp Segmentation

Shivam Gupta
Independent Researcher, Delhi, India
shivam1720406@gmail.com | ORCID 0009-0009-4576-0138

Extended abstract for NBER Applications of Artificial Intelligence in Healthcare, 2-3 June 2027. Requested format: nonarchival poster.

Acknowledgment and disclosure. No external research funding was received for this study. The author is founder of Siloed, an AI software consultancy. No Siloed product is evaluated.

Question and contribution

Recall calibration applies to the mask ultimately returned by a segmentation pipeline. Retaining only the largest connected component can delete annotated pixels after calibration, and component switches can make the cleaned family nonmonotone across thresholds. We audit these mechanisms separately. The study applies established conformal risk-control and connected-filter principles [1,2], provides a minimal running-union completion on a fixed threshold grid, and compares its utility with existing alternatives. It does not propose a new general conformal theorem.

Experimental design

Fixed PraNet and Polyp-PVT predictions are evaluated on 701 distinct images from four public polyp-segmentation datasets after removing 37 exact repeats. Seven procedures are compared over 500 shared half-pool calibration/evaluation splits. Outcomes are image-averaged missed annotated area, predicted frame area and Dice. Models and postprocessors are not refitted for each split. Mathematical invariants, connectivity choices and threshold-grid sensitivity are checked separately from the benchmark analysis.

PraNet: selected mean missed annotated area (%) at the 10% target, over the same 500 splits.

Procedure	Mean missed area
Raw conformal risk control	9.36%
Largest-component cleanup after raw calibration	48.64%
Running-union persistence completion + calibration	8.99%
Established loss-envelope calibration	9.16%
Established anchor filter + calibration	9.01%

Replication

For Polyp-PVT, cleanup increases mean missed area from 9.75% to 13.53%; persistence yields 9.78%. The model-specific difference shows why final-mask validation is necessary even when the same cleanup rule is used.

What the comparisons establish

Deleting pixels after calibration increases the missed-area loss of a fixed mask. Recalibrating the cleaned family introduces a separate issue: when the largest component changes, expanding the raw mask can select a different component and increase the loss. The running union makes the selected masks nested on the declared grid. Established loss-envelope calibration supplies a separate correction for nonmonotonicity. Neither correction is presented as uniformly preferable. In the primary PraNet comparison, an established anchor filter attains 9.01% missed area with 28.01% mean frame area, compared with 8.99% and 39.43% for persistence; it also outperforms persistence on mean Dice. The negative control is essential to interpreting the proposed completion as one available construction rather than a universally superior segmenter.

Statistical and clinical interpretation

The guarantee concerns expected image-level missed annotated area under exchangeability and the stated finite-grid assumptions. It is not a per-patient detection guarantee or evidence of complete resection. The 500 splits are paired reuse of a fixed image pool, not independent cohorts. Source-held-out stress tests can exceed the target under distribution shift, which lies outside the exchangeability argument. The public releases do not provide patient or sequence identifiers, limiting assessment of dependence across frames. These qualifications are material to the use of calibration in medical AI validation.

Reproducibility and practical use

The public artifact records upstream revisions, checksums, duplicate exclusions, per-image threshold paths and fixed split assignments. Saved missed-area, frame-area and Dice statistics support reproducing comparisons without rerunning neural inference. The manuscript reports exact reproduction of all four saved statistical analyses. An exhaustive 2-by-3-grid test checks 729 score images under both four- and eight-connectivity, and separate finite-population checks verify the calibration inequality. The practical recommendation is to validate the returned mask, quantify cleanup-induced deletion and compare repaired procedures with standard increasing filters. The poster will present the mechanism, paired evidence and useful counterexamples, with no claim of a clinical deployment study.

Availability and presentation status

Code, protocol and sufficient statistics: https://github.com/shi1720/polyp-component-persistence. Images and weights remain with their original releases and retain their original terms. An IMKASID 2027 abstract has been submitted. This request is for a nonarchival poster and makes no claim of prior acceptance or journal publication.

Selected references

[1] Angelopoulos, A. N., Bates, S., Fisch, A., Lei, L., and Schuster, T. Conformal risk control. ICLR (2024). https://arxiv.org/abs/2208.02814.

[2] Dijkstra, W., Sobiecki, A., Bernal, J., and Telea, A. C. Towards a single solution for polyp detection, localization and segmentation in colonoscopy images. VISIGRAPP, 616-625 (2019). doi:10.5220/0007694906160625.

[3] Fan, D.-P. et al. PraNet: Parallel reverse attention network for polyp segmentation. MICCAI, 263-273 (2020). doi:10.1007/978-3-030-59725-2_26.