
This MICCAI paper is the Open Access version, provided by the MICCAI Society. It is identical to the accepted version, except for the format and this watermark; the final published version is available on SpringerLink.

A Scientific Supplement

This supplement records the cohort construction, proposal mechanisms, simulated review process, declared score, and supporting results used in the main paper. All tables report frozen study artifacts; no additional model selection was performed for the supplement. The frozen configurations, the retained per-case traces, and a verifier that reproduces the reported values without ISIC images or model weights are available in the reproduction package at `https://github.com/flint-xf-fan/AuthEval`.

A.1 Scope and Limitations

AuthEval is evaluated in a controlled study built from one public dermoscopy collection. Proposals come from frozen models scored on real images, while local review, private context, and the scoring weights follow a study-defined rule. The evidence therefore characterizes an evaluation construct rather than clinical prevalence, benefit, or harm. ISIC records no referral, capacity, clinician-selection, completion, or outcome fields, so the workflow surrounding each image is authored rather than observed.

The four workflow profiles are study assignments rather than institutions, and the private factors are simulated independently of any recorded patient attribute. The local review rules are stress tests chosen to span admissible behavior, not clinical prescriptions, and the declared score is a static study score rather than an outcome model. A gap measured under one such rule constrains neither the gap under a different rule nor the clinical consequence of the substitutions producing it.

The authority-gap endpoint and mechanism analyses reported here were conducted post hoc after a prospectively locked policy-selection comparison yielded the same policy under proposal-only and authority-aware objectives.

The three proposal mechanisms differ in architecture, readout, input signals, and slate construction, so ordering them by authority gap does not isolate a single component: three arms establish that the gap moves, not why it moves. Image-equal weighting was designated primary after exploratory development, and the cluster-equal estimand accompanies it throughout because the sign of the primary estimate depends on that choice.

Stability after excluding the broader threshold-4 pHash components remains untested because that heuristic can also connect visually similar but distinct lesions.

Earlier ISIC releases entered MedGemma pretraining, so “held out” describes the fitting of the action readout and the control thresholds rather than the foundation model itself. Every estimate is conditional on the fitted readouts and on the frozen proxy, gate, score, and simulated context, and the confidence intervals propagate sampling variation across clusters alone.

Table 2. Split sizes and risk-stratum composition. Strata are derived from the recorded diagnosis and are used only as the post-hoc reference state.

Split	n	Very low	Low	Moderate	High
Train	17,629	9,392	1,839	2,890	3,508
Validation	3,735	1,923	393	648	771
Test	3,967	2,052	392	652	871

Table 3. Perceptual-duplicate groupings over all 25,331 images. Exact groups constrain the split; threshold-4 components are diagnostic only.

Grouping	Components	Images
Exact pHash	95	191
Exact pHash crossing splits	0	0
Threshold-4 pHash	594	1,777
Threshold-4 pHash crossing splits	122	—

A.2 Cohorts, Splits, and Evaluation Units

Risk strata and routing actions. The study uses the ISIC 2019 training metadata, ground truth, and image release. Diagnoses map to four ordinal risk levels: very low (melanocytic nevus, vascular lesion, or dermatofibroma), low (benign keratosis), moderate (actinic keratosis or intraepithelial carcinoma, and basal-cell carcinoma), and high (squamous-cell carcinoma or melanoma). Four ordinal routing actions $a_0, \ldots, a_3$ span observation, routine review, specialist assessment, and urgent specialist assessment. The reference state y is the risk level implied by the recorded diagnosis and enters scoring only. Table 2 reports the realized split composition.

Split construction and duplicate control. Images sharing a known lesion identifier or an exact 64-bit discrete cosine transform (DCT) perceptual hash (pHash) are assigned to the same split, so neither relation crosses the train, validation, and test boundaries. A broader pHash diagnostic at Hamming threshold 4 is computed for reference but does not redefine the split, because that threshold also connects visually similar yet distinct lesions. Table 3 summarizes both groupings. The 122 components that cross splits under the broader threshold are the reason the diagnostic is reported rather than enforced.

Evaluation units. The primary estimand weights the 3,967 test images equally. The sensitivity estimand weights 2,090 clusters equally, of which 1,778 derive from known lesion identifiers and 312 from images carrying none, after exact-hash unions. Each cluster receives one simulated context, so its members share a workflow profile, private factors, and mean metadata proxy score. Confidence intervals resample whole clusters, using 5,000 replicates and the 0.025 and 0.975 quantiles.

Table 4. Test-split cluster sizes by risk stratum. Image-equal weighting is proportional to the image column, cluster-equal weighting to the cluster column.

Risk stratum	Clusters	Images	Images per cluster
Very low	1,333	2,052	1.54
Low	221	392	1.77
Moderate	244	652	2.67
High	292	871	2.98

The two units are not interchangeable, because lesions are photographed unequally across the risk strata. Table 4 gives the imbalance: high-risk clusters carry roughly twice as many images as very-low-risk clusters, so image-equal weighting places proportionally more weight on high-risk lesions than cluster-equal weighting does. This is the mechanism behind the sign reversal reported in the main paper.

A.3 Proposal Mechanisms

MedGemma proposal. We run the primary proposer, MedGemma 1.5 4B IT, in bfloat16 at the frozen identifier and revision:

`google/medgemma-1.5-4b-it`
`91850547d9f0b2fdd21aa7c5f4f3d1a8a52c243b`

For each image and the proposer-visible age, sex, and anatomic site, the frozen model assigns a mean token log-likelihood to four semantic continuations under the prompt:

> This is a research image-assessment task, not patient-specific clinical advice. Inspect the dermoscopy image and the listed public metadata only. Rank the lesion into one of four ordinal visual-risk categories. Do not infer unavailable history, symptoms, evolution, patient preference, local policy, or service capacity. Categories: very low visual risk; low visual risk; moderate visual risk; high visual risk. Public metadata:
> `age={age}, sex={sex},`
> `anatomic_site={anatomic_site}`. Selected visual-risk category:

The four continuations describe, respectively, a benign appearance without evident suspicious dermoscopic features; a mostly benign appearance with limited atypical features; suspicious features warranting specialist assessment; and strongly suspicious features warranting urgent specialist assessment.

Action readout. A balanced multinomial logistic readout maps the four log-clipped continuation probabilities (clip 10^{-8}) to the four routing actions. Four-fold grouped validation searches $C \in \{0.01, 0.1, 1, 10\}$, minimizing ordinal mean absolute error and then maximizing balanced accuracy and macro-F1, with smaller C as the final tie-break; the selected value is $C = 0.01$. Test labels

enter neither fitting nor selection. On the held-out validation set, the readout uses all four actions and improves ordinal MAE from 1.104 for the best constant to 0.860. Shuffling images against their metadata within a 256-image validation subset raises MAE by 0.375 and lowers balanced accuracy by 0.132, indicating that the readout responds to image content rather than metadata alone.

Proposal controls. Two frozen controls supply alternative proposals through the same interface:

`microsoft/llava-med-v1.5-mistral-7b`
`openai/clip-vit-large-patch14-336`

The language control reads option likelihoods from the first and converts its expected ordinal score to an action; the vision control uses a balanced linear probe trained on frozen CLIP ViT-L/14 features from the second. Both discretize with validation-fitted thresholds:

$$\begin{aligned} \text{language control:} &\quad (0.602433,\ 0.645141,\ 0.660500), \\ \text{vision control:} &\quad (0.459906,\ 0.551713,\ 0.662236). \end{aligned}$$

Both controls are consumed by the simulator only through these frozen operational heads, each a function of the image and declared public metadata alone.

A.4 Simulated Review Process

Workflow profiles. Each cluster is assigned one of four workflow profiles by a stable hash of its age band and anatomic site, falling back to the image identifier only when both fields are missing; a cluster takes the modal profile of its members. A profile fixes an upper capacity bound on escalation, a floor policy, and the base under- and over-escalation weights of Table 5. The profiles are labeled by the service character they are meant to span, from a high-capacity dermatology service to a low-capacity, access-constrained one.

Table 5. Workflow-profile parameters. Capacity is the highest admissible action level; w_u and w_o are the base under- and over-escalation weights before the private multipliers of Table 6.

Profile	Service character	Capacity	Floor policy	w_u	w_o
P1	High-capacity dermatology	3	Standard	1.20	0.75
P2	General mixed source	2	Standard	1.20	1.20
P3	Specialist, safety-oriented	3	Strict	1.80	0.90
P4	Low-capacity, access-constrained	1	Lenient	1.30	2.00

Table 6. Private factor levels and their multipliers on the under- and over-escalation weights.

Factor	Level	w_u mult.	w_o mult.
Sensitivity preference	Sensitivity-seeking	1.25	0.95
	Neutral	1.00	1.00
	Specificity-seeking	0.95	1.25
Follow-up burden	Low	1.00	1.00
	High	1.00	1.15
Invasiveness aversion	Neutral	1.00	1.00
	Averse	1.00	1.20

Private factors. Within each cluster, one level of each factor in Table 6 is drawn and held fixed. Sensitivity preference rescales both weights; follow-up burden and invasiveness aversion rescale the over-escalation weight alone. Invasiveness aversion additionally removes a_3 from the admissible set, and high follow-up burden replaces an admissible a_1 by a_0 in lower proxy bins or by a_2 in higher bins when that alternative is itself admissible.

Gate proxy and selection rule. The decision-time risk context is a class-balanced ridge model fitted on training labels from age, sex, and anatomic site, with regularization and temperature selected on validation. Its score is binned at $(0.25, 0.50, 0.75)$. The profile supplies the capacity bound, the proxy bin together with the profile floor policy supplies the safety floor, and the mixed regime intersects both. When an override empties the admissible band, the declared default is a_0 in lower proxy bins and a_2 in higher bins. The gate selects the first admissible action in the ranked slate and never observes the diagnosis, the reference action, or the score, so selection cannot depend on the evaluation target.

A.5 Declared Score

Score definition. For action and reference levels $\ell(a), \ell(y) \in \{0, 1, 2, 3\}$, the declared score is

$$U_z(x, y, a) = -w_u(z)\, s_y\, [\ell(y) - \ell(a)]_+ - w_o(z)\, [\ell(a) - \ell(y)]_+, \tag{1}$$

with $[t]_+ = \max(t, 0)$ and s_y the component indexed by $\ell(y)$ in $(s_0, s_1, s_2, s_3) = (0.5, 1, 2, 4)$. Zero is the best attainable value.

Weight ranges and asymmetry. Applying the multipliers of Table 6 to the base weights of Table 5 gives effective ranges $w_u \in [1.14, 2.25]$ and $w_o \in [0.71, 3.45]$. A one-level over-escalation therefore costs between 0.71 and 3.45 units, whereas a one-level under-escalation at high risk costs between 4.6 and 9.0, since $s_3 = 4$ multiplies the under-escalation weight. Upward and downward substitutions are

consequently not equivalent under Eq. (1), contributing to the different gap magnitudes and signs produced by the two review processes.

Relation to a benchmark metric. A benchmark may report a score J_{bench} under a different metric. When J_{bench} is commensurate with U_z and uses the same cases, evaluation distribution, and weighting rule, its difference from the selected-action score decomposes as

$$J_{\text{bench}} - J = (J_{\text{bench}} - J_{\text{auth}}) + \Delta_{\text{authority}}, \tag{2}$$

where the first term is the gap between the benchmark metric and the declared local score evaluated on the same proposal, and the second is the local-selection effect isolated in the main paper. A comparison that does not hold the first term fixed cannot attribute a change in $J_{\text{bench}} - J$ to local selection alone.

A.6 Record Levels

Table 7 states which AuthEval quantities each cumulative record level supports. The levels are cumulative: each adds the fields shown to the preceding level and thereby supports the quantities listed without additional assumptions.

Every record also versions the case and cluster identifiers, proposal interface, local selection and scoring rules, reference state, and evaluation unit, and may log actor, reason, timestamp, and completion as further workflow metadata. If $p = 0$, AuthEval reports $\Delta_{\text{authority}} = 0$ and leaves δ undefined; if $p > 0$ but the declined proposal's local score is unavailable, it reports $\Delta_{\text{authority}}$ and δ as not recoverable. Exporting locally computed scores rather than raw context can reduce data transfer, but provides no formal privacy guarantee.

Table 7. Cumulative record levels and the AuthEval quantities each one recovers. Fields are additional to the level above.

Record level	Additional local-review fields	Quantities recovered
Proposal-only	$\hat{a}$ and proposal metric	none
Role-separated	a_{sel}, $U_z(x, y, a_{\text{sel}})$	p, J
Full counterfactual	$U_z(x, y, \hat{a})$ on changed cases	J_{auth}, $\Delta_{\text{authority}}$, δ if $p > 0$

A.7 Supporting Results

Review-process decomposition. Table 8 reports the two constituent review processes alongside the mixed process of the main paper. The proposal score J_{auth} is identical across the three rows because the proposal does not depend on the gate; only the selected-action score J moves. The capacity ceiling constrains actions through the private profile and factors, and the safety floor constrains them

Table 8. Authority gap by simulated review process for the fixed MedGemma-derived proposal. Scores are in simulator units; brackets give 95% whole-cluster bootstrap intervals.

Review process	J_{auth}	J	p	δ	$\Delta_{\text{authority}}$, image	$\Delta_{\text{authority}}$, cluster
Capacity ceiling	-2.938	-3.683	0.341	2.182	$+0.744$ $[+0.491, +1.010]$	$+0.056$ $[-0.107, +0.218]$
Safety floor	-2.938	-2.888	0.281	-0.180	-0.051 $[-0.262, +0.158]$	-0.224 $[-0.358, -0.093]$
Mixed	-2.938	-3.313	0.470	0.797	$+0.374$ $[+0.091, +0.663]$	-0.206 $[-0.373, -0.037]$

through the public metadata proxy that the proposer also sees. Under image weighting the discrepancy is carried by the capacity channel, whose substitutions are 97.6% downward, while the safety channel is indistinguishable from zero; under cluster weighting the ordering reverses, and the safety channel is the one whose interval excludes zero.

Proposal dependence. Table 9 holds the capacity gate, cases, private context, clusters, and score fixed while varying the proposal mechanism. Both the change rate and the conditional score change vary with the mechanism; under cluster weighting, the language control and vision probe have authority gaps of opposite sign. The paired contrast evaluates both proposals on the same clusters within each replicate, and its confidence interval excludes zero under both evaluation units.

Table 9. Authority gap by proposal mechanism under one fixed capacity review process. The final row is the paired probe-minus-LLaVA-Med contrast, computed on identical clusters within each bootstrap replicate.

Proposal	Test balanced acc.	p	δ	$\Delta_{\text{authority}}$, image	$\Delta_{\text{authority}}$, cluster
LLaVA-Med head	0.249	0.163	0.084	$+0.014$ $[-0.056, +0.086]$	-0.096 $[-0.156, -0.037]$
MedGemma readout	0.420	0.341	2.182	$+0.744$ $[+0.491, +1.010]$	$+0.056$ $[-0.107, +0.218]$
CLIP probe	0.546	0.276	2.988	$+0.824$ $[+0.621, +1.033]$	$+0.337$ $[+0.256, +0.422]$
Probe − LLaVA-Med	—	—	—	$+0.810$ $[+0.629, +0.999]$	$+0.433$ $[+0.345, +0.524]$