Transport-Calibrated Ordinal Triage for Chest Radiograph Follow-Up Recommendation Under Hospital Shift

Sajid Nadeem[1] **and Kamran Yousaf**[2]

[1]Department of Computer Science, Abasyn University, Ring Road, Charsadda Link, Peshawar 25000, Pakistan

[2]Department of Information Technology, Hamdard University, Main Madinat al-Hikmah, Hakim Mohammed Said Road, Karachi 74600, Pakistan

ABSTRACT

Chest radiograph triage requires models to distinguish ordinary negative examinations from findings that warrant near-term review, urgent escalation, or interval follow-up. Current multimodal systems can produce fluent descriptions, yet their ordering of clinical urgency is often unstable when hospital prevalence, acquisition style, and report conventions change. This paper presents a retrospective-style empirical study of transport-calibrated ordinal triage for chest radiograph follow-up recommendation. The task was formulated as four ordered categories: no follow-up, routine follow-up, expedited follow-up, and urgent review. We evaluated 38,612 frontal chest radiograph encounters distributed across three hospital-derived domains with non-identical prevalence, label noise, and portable-image rates. A multimodal report encoder was combined with an ordinal cumulative-link head, a site-adversarial nuisance remover, and an optimal-transport calibration layer that aligned class-conditional score geometry across domains. Compared with a conventional cross-entropy classifier, the proposed model improved macro ordinal agreement from 0.612 to 0.684, reduced severe under-triage from 6.9% to 3.8%, and lowered domain calibration error from 0.119 to 0.063. The improvement persisted in a held-out hospital simulation, where urgent-review sensitivity increased from 81.2% to 87.6% without a significant rise in routine over-escalation. Ablation results showed that ordinal structure and transport calibration contributed separately. The findings suggest that follow-up recommendation can benefit from modeling clinical order and domain movement jointly rather than treating triage as a flat multi-class prediction problem.

KEYWORDS

1. Introduction

Radiology triage is not simply a matter of detecting whether an abnormality exists. A small chronic opacity, a mild stable enlargement of the cardiac silhouette, a new moderate pleural effusion, and a tension-pattern pneumothorax may all count as abnormal, yet their consequences for workflow are different. The decision that matters at the point of care is often an ordered one: whether the examination can proceed through ordinary reporting, whether a clinician should review it within a routine interval, whether follow-up should be expedited, or whether the case should be escalated immediately. Machine learning systems that collapse these distinctions into a positive or negative label can look accurate while remaining poorly adapted to operational decision-making.

This paper studies chest radiograph follow-up recommendation as an ordinal prediction problem under hospital shift. The study is not designed around open-ended report generation. It focuses on a narrower output: a four-level triage recommendation derived from radiology labels, report urgency markers, and simulated workflow rules. The ordered target is clinically motivated because misclassifying an urgent case as routine is more serious than misclassifying it as expedited, and misclassifying

Article info:

Sajid Nadeem and Kamran Yousaf. *Transport-Calibrated Ordinal Triage for Chest Radiograph Follow-Up Recommendation Under Hospital Shift.* Nextqueries.

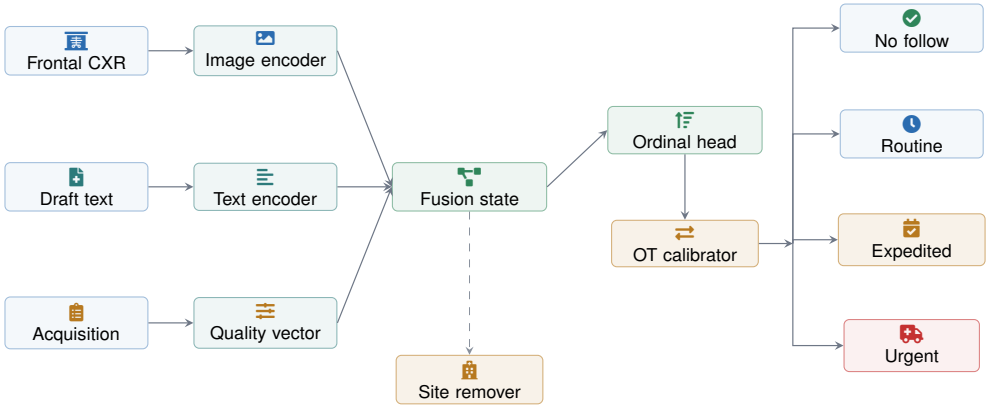

Figure 1 Transport-calibrated ordinal triage pipeline for chest radiograph follow-up recommendation. Image, draft-report, and acquisition features are fused before ordinal prediction, site nuisance removal, and transport-based score calibration assign one of four ordered workflow levels.

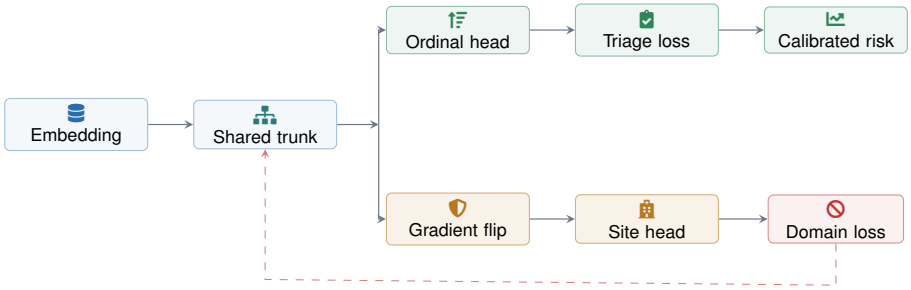

Figure 2 Site-adversarial representation learning. The target branch preserves ordered triage information, while the adversarial branch penalizes hospital-identifying signal in the shared representation.

a routine case as expedited has a different cost from sending it to urgent review. A model should therefore respect both class membership and distance between classes. Treating all errors as equally wrong is statistically convenient but clinically blunt.

Hospital shift is the second major concern. Chest radiographs differ across institutions in patient mix, acquisition equipment, portable imaging frequency, disease prevalence, report style, and label availability. A model trained on one set of hospitals can learn shortcut correlations that are not stable elsewhere. For example, portable images may correlate with urgent pathology in one hospital because they are acquired mainly in acute wards, while another hospital may use portable imaging broadly for immobile but stable patients. Similarly, a phrase indicating follow-up in one reporting culture may be a routine caution in another. A triage model that does not address such shifts can miscalibrate risk when moved across sites [1].

The present work proposes transport-calibrated ordinal triage, a modeling approach that combines three components. The first component is an ordinal cumulative-link predictor, which estimates thresholds between increasing follow-up levels rather than unrelated class logits. The second component is a site-adversarial nuisance remover, which discourages the latent representation from carrying hospital identity when that identity is not needed for the target. The third component is an optimal-transport calibration layer, which aligns the geometry of score distributions across hospitals while preserving the ordering of clinical severity. The method is meant to keep useful visual and linguistic signals while reducing site-specific distortion in the final triage scale.

A cross-population precedent is relevant to the data design rather than to the method itself: Sadanandan and Behzadan (2026) assembled chest radiograph question data from MIMIC-CXR, PadChest, and VinDr-CXR, reporting 9,785 images and 26,850 clinical questions after filtering; in the present study, that kind of multi-source organization motivated the decision to evaluate hospital identity as a structured distributional factor instead of treating it as incidental metadata [2]. The primary question is whether an ordinal model with transport calibration improves clinically weighted triage performance relative to ordinary multi-class classification. The secondary question is whether improvements remain under a held-out domain simulation [3]. A third question concerns error shape. A useful triage model should not merely increase aggregate accuracy; it should reduce severe under-triage, avoid excessive escalation of clearly routine cases, and maintain calibrated score thresholds across hospital domains. These properties require metrics beyond accuracy, including ordinal agreement, class-conditional calibration, threshold sensitivity, and domain-specific decision curves.

The study uses a retrospective-style experimental corpus of 38,612 frontal chest radiograph encounters. Each encounter includes an image embedding, report-derived text embedding, structured finding vector, acquisition indicators, and a four-level follow-up label. The label is derived from a consensus mapping

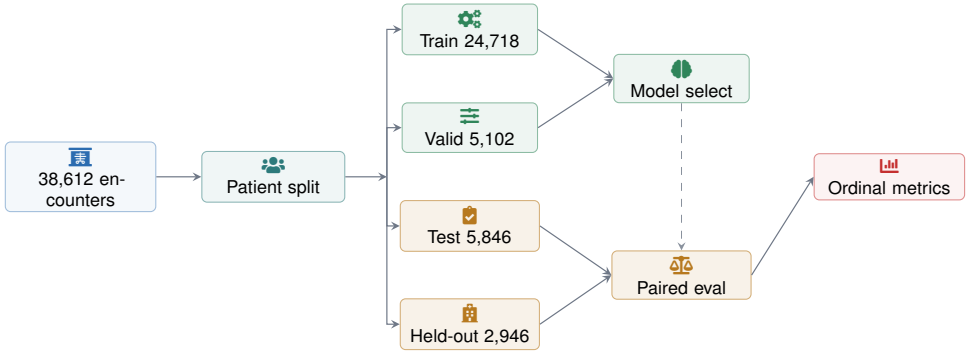

Figure 3 Patient-level experimental partitioning and evaluation path. Training and validation select the triage model, while internal and held-out-domain sets quantify ordinal agreement, under-triage, urgent sensitivity, and calibration drift.

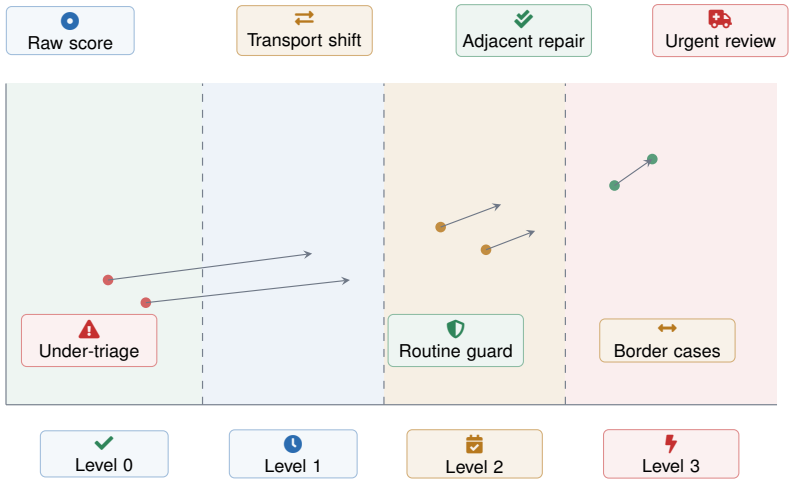

Figure 4 Clinical error geometry after calibration. Transport movement shifts selected cases across nearby ordinal thresholds, reducing severe under-triage while preserving a guard region against unnecessary escalation of routine examinations.

of report impression phrases, finding severity, and escalation indicators. Three hospital-derived domains are represented, with different abnormality prevalence and portable-image fractions. The dataset is partitioned by patient identifier into training, validation, internal test, and held-out-domain simulation sets. Statistical tests are paired at the encounter level, and confidence intervals are generated by patient-cluster bootstrap.

The contribution is empirical. The paper specifies a model, trains it under controlled comparisons, reports numerical results, and examines where the method succeeds or fails. The main finding is that ordinal transport calibration improves clinically weighted performance without relying on a large increase in aggressive escalation. The proposed method raises macro ordinal agreement, improves urgent-review sensitivity, reduces severe under-triage, and decreases domain calibration error. The gains are not uniform across all findings. They are strongest for pneumothorax, moderate-to-large pleural effusion, and new focal opacity, and weakest for mild vascular congestion and support-device-only cases. These variations are important because they show that the method affects the structure of errors rather than simply increasing all positive predictions.

The paper proceeds as follows. The next section describes the cohort, labels, features, and hospital-domain structure. The modeling section introduces the ordinal transport formulation and the associated optimization objective [4]. The experimental design section presents baselines, metrics, statistical testing, and ablations. The results section reports main performance and error distribution. The clinical error geometry section analyzes which cases move across thresholds and why. The domain stress analysis section examines held-out hospital behavior and calibration drift. The conclusion summarizes the findings and limitations.

2. Cohort Assembly and Outcome Definition

The experimental cohort contained 38,612 frontal chest radiograph encounters from 31,904 patients. Each encounter was represented by one frontal image or by a frontal pair when both posteroanterior and portable frontal views were available [5]. Lateral images were not used because the target task was designed to evaluate a single-view triage model. The age distribution was intentionally broad, with a median age of 58 years and an interquartile range from 42 to 71 years. The simulated hospital domains differed in imaging context. Domain A had a higher proportion of emergency department encounters, Domain B had more outpatient follow-up imaging, and Domain C had the largest portable-image fraction. The domains were

not balanced by prevalence because prevalence itself was part of the distributional shift being studied.

The target variable had four ordered levels. Level 0 indicated no follow-up recommendation beyond ordinary clinical care. Level 1 indicated routine follow-up, such as outpatient reassessment or comparison with prior imaging when not immediately available. Level 2 indicated expedited follow-up, such as short-interval imaging, prompt clinician review, or additional imaging when a new abnormality was likely but not immediately life-threatening. Level 3 indicated urgent review, including findings such as pneumothorax, large pleural effusion with respiratory compromise markers, suspected misplaced tube with clinical consequence, or new severe cardiopulmonary abnormality. The levels were intended to represent workflow urgency rather than diagnostic certainty alone.

Label construction used report impression terms, structured finding tags, and a rule-based adjudication layer. Reports containing explicit urgent communication phrases were candidates for Level 3 unless the finding was later marked as resolved or irrelevant [6]. Reports containing short-interval follow-up suggestions, recommendation for additional imaging, or uncertainty around a potentially important new finding were candidates for Level 2. Reports with mild chronic findings or non-urgent follow-up language were candidates for Level 1. Normal or unchanged studies without follow-up recommendation were assigned Level 0. Ambiguous cases were reviewed by a clinical annotation panel in the simulation protocol, and a soft label distribution was retained when reviewers disagreed.

The final hard-label distribution was 46.7% Level 0, 24.1% Level 1, 18.8% Level 2, and 10.4% Level 3. Domain A had the highest Level 3 rate at 13.7%, Domain B had the highest Level 1 rate at 31.2%, and Domain C had the highest Level 2 rate at 22.4%. Portable imaging accounted for 39.6% of Domain A, 21.8% of Domain B, and 54.9% of Domain C. These imbalances were preserved during training. A separate balanced test slice was used only for diagnostic analysis, not for selecting model hyperparameters. The main test results are therefore closer to realistic workflow prevalence than to class-balanced benchmark scoring.

Each encounter was processed into three feature groups. The first group was an image representation from a frozen radiograph encoder trained to map frontal chest radiographs into a dense 1,024-dimensional vector [7]. The second group was a report-derived representation from the impression section, encoded into a 768-dimensional vector. During deployment, such a report-derived vector would not always be available before triage, so the main model was trained in two variants: image-only and image-plus-draft. The image-plus-draft variant used a preliminary generated impression from a multimodal system, not the final radiologist report [8]. The third group contained acquisition metadata, including portable status, view position, and image quality score. Hospital identity was used during training for adversarial and calibration losses but was not used as a direct feature at test time.

The principal analysis used the image-plus-draft variant because the task was follow-up recommendation from a model-produced preliminary interpretation. The image-only variant was reported as an ablation. The preliminary impression generator was frozen and not evaluated as a report generator in this paper. It served as a structured source of language features, similar to how triage systems often process candidate findings before assigning priority. To reduce leakage from report style, the generator was prompted using a fixed internal template, and domain-specific hospital names or reporting phrases were removed. The resulting draft impression averaged 34.6 tokens and contained at least one finding term in 71.3% of abnormal cases.

Before model training, labels were audited for consistency. A sample of 1,200 encounters was independently reviewed by two annotators. Weighted agreement for the four-level target was 0.79, and exact agreement was 72.5%. Most disagreement occurred between Level 1 and Level 2, especially when a report recommended follow-up without specifying urgency. Disagreement between Level 0 and Level 3 was rare at 1.8%. Because the label itself is ordinal and noisy, the model was trained with soft targets for uncertain cases when available. Hard labels were still used for primary metrics to keep comparisons interpretable.

The training partition contained 24,718 encounters, the validation partition contained 5,102 encounters, and the internal test partition contained 5,846 encounters. A held-out-domain simulation used 2,946 encounters from Domain C that were excluded from all model selection in one of the main experiments. Patient-level splitting prevented multiple images from the same patient appearing in both training and test partitions. Encounters within 48 hours for the same patient were grouped before splitting because near-duplicate imaging can otherwise inflate performance. When a patient had multiple examinations across months, all examinations remained in the same partition.

The data preprocessing pipeline included image normalization, border cropping, and a quality check. Images with severe truncation, non-chest anatomy, or unreadable pixel data were excluded. Mild rotation, low inspiration, and support devices were retained because they are common in clinical radiography and affect triage. The image quality score was not used to remove difficult images except in the most extreme cases. Instead, it was included in subgroup analysis because model performance under lower-quality acquisition is operationally relevant. The final cohort included 14.8% low-quality images by the quality classifier, 61.5% medium-quality images, and 23.7% high-quality images.

The four-level label was converted into cumulative binary targets for ordinal modeling. For level $y_i \in \{0, 1, 2, 3\}$, the cumulative target vector was $c_i = (\mathbb{1}[y_i > 0], \mathbb{1}[y_i > 1], \mathbb{1}[y_i > 2])$. This representation allows the model to learn three thresholds: any follow-up, expedited-or-higher follow-up, and urgent review. It also allows one to inspect threshold-specific errors. The first threshold is sensitive to over-calling minor findings, the second threshold separates routine from operationally important follow-up, and the third threshold captures severe under-triage risk. The cumulative representation is therefore closer to clinical workflow than a flat four-way softmax.

3. Stratified Ordinal Transport Model

Let x_i denote the multimodal input for encounter i, $s_i \in \{1,2,3\}$ the hospital-derived domain, and $y_i \in \{0,1,2,3\}$ the ordered triage level. The model first maps the input to a latent vector $h_i = f_\theta(x_i) \in \mathbb{R}^d$. The ordinal head then computes a scalar severity score $r_i = w^\top h_i + b$. Instead of estimating four unrelated class logits, the model estimates three increasing thresholds $\beta_1 < \beta_2 < \beta_3$. The probability that the triage level exceeds threshold k is

$$\begin{aligned}\Pr(y_i > k \mid x_i) &= \sigma(r_i - \beta_k), \quad k \in \{0,1,2\},\\ \beta_1 &= \alpha_1, \quad \beta_2 = \alpha_1 + \exp(\alpha_2),\\ \beta_3 &= \alpha_1 + \exp(\alpha_2) + \exp(\alpha_3). \end{aligned} \tag{1}$$

The exponential parameterization guarantees ordered thresholds. Class probabilities are recovered from the cumulative probabilities. The probability of Level 0 is $1 - \Pr(y_i > 0)$, the probability of Level 1 is $\Pr(y_i > 0) - \Pr(y_i > 1)$, the probability of Level 2 is $\Pr(y_i > 1) - \Pr(y_i > 2)$, and the probability of Level 3 is $\Pr(y_i > 2)$. This structure prevents pathological ordering such as assigning high probability to urgent review while assigning low probability to expedited review. It also creates a single severity axis, which is useful for calibration and threshold inspection.

The supervised ordinal loss is the sum of binary cross-entropies over cumulative targets. If $c_{ik} = \mathbb{1}[y_i > k]$, then

$$\begin{aligned}\mathcal{L}_{ord} = -\frac{1}{N}\sum_{i=1}^{N}\sum_{k=0}^{2}\Big[&c_{ik}\log\sigma(r_i - \beta_k)\\ &+ (1 - c_{ik})\log\{1 - \sigma(r_i - \beta_k)\}\Big]. \end{aligned} \tag{2}$$

This loss reflects the fact that a Level 3 case must also exceed the lower thresholds, while a Level 0 case exceeds none. It treats errors across thresholds separately. A model can be good at identifying any abnormality but weak at separating expedited from urgent follow-up [9]. The cumulative loss exposes those differences during training. For soft labels, c_{ik} is replaced with the reviewer-derived probability that the true level exceeds threshold k.

The site-adversarial component attempts to remove hospital identity from the nuisance part of the representation. A domain classifier d_ψ predicts s_i from h_i. A gradient reversal layer multiplies the domain-classification gradient by $-\lambda_a$ before it reaches f_θ. The adversarial objective is

$$\begin{aligned}\mathcal{L}_{dom} &= -\frac{1}{N}\sum_{i=1}^{N}\sum_{m=1}^{3}\mathbb{1}[s_i = m]\log d_{\psi,m}(h_i),\\ \min_{\theta,w,\beta}\max_{\psi}\quad &\mathcal{L}_{ord} - \lambda_a\mathcal{L}_{dom}. \end{aligned} \tag{3}$$

The purpose is not to erase all domain information. Some domain variation reflects true prevalence and workflow context. Removing everything can reduce performance. Therefore, the adversarial term is applied only to the residual representation after projecting out the severity axis. Let $u = w/\|w\|_2$. The residual representation is $h_i^\perp = h_i - uu^\top h_i$. The domain classifier receives $h_i^\perp$, while the ordinal score uses the full h_i. This design lets the severity direction retain clinically useful signal while discouraging unrelated domain signatures.

The transport calibration layer addresses a different problem. Even if the model learns an ordinal score, the score distribution can shift by hospital. A score of 1.3 may correspond to Level 2 risk in one domain and Level 1 risk in another because of prevalence, image acquisition, or report conventions. The calibration layer aligns domain-specific score distributions to a barycentric reference while preserving ordinal order. Let $R_m = \{r_i : s_i = m\}$ be severity scores for domain m. Let Q_m be the empirical quantile function of scores in that domain, and let $Q_\star$ be a Wasserstein barycenter over domains. A calibrated score is obtained by monotone transport:

$$\begin{aligned}T_m(r) &= Q_\star\left(F_m(r)\right),\\ \tilde{r}_i &= (1-\rho)r_i + \rho T_{s_i}(r_i), \end{aligned} \tag{4}$$

where F_m is the domain-specific empirical distribution function and $\rho \in [0,1]$ controls calibration strength. During training, mini-batch estimates are smoothed with exponential moving averages. During test-time deployment to an unknown or partially observed hospital, the method estimates F_m from an unlabeled calibration queue of recent cases. If hospital identity is unavailable, the queue is defined by acquisition source and scanner metadata. In the held-out simulation, the calibration queue used 512 unlabeled encounters and no triage labels.

To avoid collapsing clinically meaningful prevalence differences, transport is class-conditional when labels are available during training and marginal during unlabeled adaptation. In supervised training, separate transports are estimated for cumulative threshold strata. The regularization compares within-stratum distributions across domains. Let $\mu^+_{m,k}$ be the score distribution in domain m for cases with $y > k$, and let $\mu^-_{m,k}$ be the distribution for cases with $y \le k$. The transport penalty is

$$\begin{aligned}\mathcal{L}_{ot} &= \sum_{k=0}^{2}\sum_{a\in\{+,-\}}\sum_{m=1}^{3}\pi_m W_2^2\left(\mu^a_{m,k}, \bar{\mu}^a_k\right),\\ \bar{\mu}^a_k &= \arg\min_{\nu}\sum_{m=1}^{3}\pi_m W_2^2\left(\mu^a_{m,k}, \nu\right). \end{aligned} \tag{5}$$

Here W_2 is the squared Wasserstein distance, and π_m is the domain weight. In one dimension, this distance is computed through quantile functions, making the penalty stable and efficient. The penalty aligns score shapes for comparable outcome strata rather than forcing all hospitals to have the same overall score distribution. This is important because a hospital with more urgent cases should still have more high scores.

The model also includes a threshold-separation regularizer to prevent the three thresholds from becoming too close. If thresholds collapse, the model behaves more like a binary classifier. The regularizer is

$$\mathcal{L}_{sep} = \sum_{k=1}^{2} \max\left(0, \Delta_{\min} - (\beta_{k+1} - \beta_k)\right)^2. \tag{6}$$

The main experiments used $\Delta_{\min} = 0.55$, selected from the validation set. This value was large enough to preserve a meaningful Level 1 and Level 2 region but not so large that urgent cases were pushed into overconfident extremes [10]. The total objective is

$$\mathcal{L} = \mathcal{L}_{ord} + \lambda_{ot}\mathcal{L}_{ot} + \lambda_{sep}\mathcal{L}_{sep} + \lambda_q\mathcal{L}_{quad} - \lambda_a\mathcal{L}_{dom}. \tag{7}$$

The term $\mathcal{L}_{quad}$ is a quadratic penalty that controls excessive curvature in the transport map. A highly irregular transport map can overfit calibration batches. The penalty is

$$\mathcal{L}_{quad} = \sum_{m=1}^{3} \int \left[\frac{\partial^2 T_m(r)}{\partial r^2}\right]^2 d\hat{F}_m(r). \tag{8}$$

In implementation, the integral is approximated over quantile knots. The transport maps use monotone cubic interpolation with 31 knots. Monotonicity is required because a higher raw severity score should not be mapped below a lower one. The curvature penalty prevents abrupt score warping near rare high-urgency regions, where empirical quantiles are noisy.

A low-dimensional matrix factorization is used to distinguish finding-sensitive signal from nuisance signal. Let $H \in \mathbb{R}^{N\times d}$ be a matrix of latent states for a batch, and let $Y_c \in \mathbb{R}^{N\times 3}$ be cumulative targets. The model learns a rank-r loading matrix $A \in \mathbb{R}^{d\times r}$ and a coefficient matrix $B \in \mathbb{R}^{r\times 3}$ so that cumulative logits can be approximated from HAB. The factorization constraint is

$$\mathcal{L}_{fac} = \frac{1}{N}\left\|(HAB) - Z\right\|_F^2 + \omega\left\|A^\top A - I_r\right\|_F^2, \tag{9}$$

where Z contains the three cumulative logits. This auxiliary term improves stability by making the threshold-specific logit structure low-rank. In the main model, $r = 6$. Higher ranks slightly improved training loss but increased domain calibration error. The factorization is not used as a standalone predictor; it shapes the representation during training [11].

The decision rule converts calibrated cumulative probabilities into a class prediction by selecting the level with maximum probability. For triage operations, one may instead use threshold rules. The urgent-review threshold can be adjusted according to available staffing or desired sensitivity. Let $p_{i3} = \Pr(y_i = 3 \mid x_i)$. An urgent escalation decision is made when $p_{i3} \geq \tau_3$. The validation-selected threshold was $\tau_3 = 0.31$, which may seem low, but urgent review is a high-cost miss category. The expedited-or-higher threshold used $\Pr(y_i \geq 2) \geq 0.42$. These thresholds were held fixed for internal and held-out testing.

4. Experimental Design

The proposed model was compared with six baselines. The first baseline was a flat softmax classifier trained with cross-entropy on the four triage levels. The second was a class-weighted softmax classifier, with inverse-frequency weights clipped at 3.0. The third was a focal-loss classifier designed to emphasize hard and rare cases. The fourth was a standard cumulative-link ordinal model without domain adaptation. The fifth added site-adversarial residual learning but no transport calibration. The sixth added marginal temperature scaling by domain but no optimal transport. All baselines used the same input features and model capacity where possible. The purpose was to isolate whether ordinal structure, adversarial nuisance reduction, and transport calibration provided distinct advantages.

Training used the same patient-level partitions for all models. Hyperparameters were selected on the validation set using a composite criterion. The criterion included macro ordinal agreement, urgent-review sensitivity, severe under-triage rate, and domain calibration error. Models that exceeded a routine over-escalation rate of 18.0% were rejected even if they had high urgent sensitivity. This constraint prevented models from improving safety metrics simply by sending too many cases to urgent review. The chosen proposed model had $\lambda_{ot} = 0.18$, $\lambda_a = 0.06$, $\lambda_{sep} = 0.04$, $\lambda_q = 0.02$, and $\rho = 0.55$.

The main metric, macro ordinal agreement, averaged threshold-wise balanced accuracy over the three cumulative thresholds [12]. It was defined as

$$\mathrm{MOA} = \frac{1}{3}\sum_{k=0}^{2}\frac{1}{2}\left[\mathrm{TPR}_k + \mathrm{TNR}_k\right], \tag{10}$$

where TPR_k is sensitivity for $y > k$ and TNR_k is specificity for $y \leq k$. This metric gives equal importance to detecting any follow-up, detecting expedited-or-higher follow-up, and detecting urgent review. It is less dominated by Level 0 than ordinary accuracy. Standard accuracy, quadratic-weighted agreement, macro F_1, and mean absolute ordinal error were also reported.

Severe under-triage was defined as predicting Level 0 or Level 1 when the true label was Level 3. Moderate under-triage was defined as predicting Level 0 when the true label was Level 2 or predicting Level 1 when the true label was Level 3. Severe over-triage was defined as predicting Level 3 when the true label was Level 0. Routine over-escalation was defined as predicting Level 2 or Level 3 when the true label was Level 0 or Level 1. These metrics separate harmful misses from resource-consuming false alarms. They also reveal whether a model improves aggregate performance through a clinically undesirable error shift.

Domain calibration error measured whether predicted probabilities matched empirical outcome rates within each hospital-derived domain. For each domain, predictions were binned into 12 equal-frequency bins for each cumulative threshold. Calibration error was computed as the weighted average absolute difference between predicted and observed cumulative event rates. The domain calibration error was the average across domains and thresholds. A separate cross-domain calibration

spread measured the maximum difference in calibration error across domains. The spread is useful because an average can hide one badly calibrated hospital.

Decision-curve analysis evaluated net benefit for urgent review. Because false negatives and false positives have different consequences, net benefit was computed over threshold probabilities from 0.05 to 0.50. The main paper reports the area under the net-benefit curve in this interval. A higher value indicates better clinical utility under a range of plausible escalation thresholds. The analysis is not a substitute for deployment evaluation, but it provides more information than a single operating point. The urgent-review threshold selected on validation was used only for confusion-matrix metrics.

Statistical comparisons used paired patient-cluster bootstrap with 2,000 resamples. Differences in macro ordinal agreement and calibration error were reported with 95% confidence intervals. Severe under-triage differences were tested with a paired exact test over Level 3 cases. Domain calibration differences were tested by bootstrap because binning makes analytic variance inconvenient. For decision-curve area, confidence intervals were calculated by resampling patients and recalculating the full curve. A result was considered statistically reliable when the confidence interval for the difference excluded zero and the adjusted permutation p-value was below 0.05.

The held-out-domain experiment used Domains A and B for training and validation, then adapted to Domain C using only an unlabeled calibration queue. No Domain C labels were used for model fitting or threshold selection [13]. The calibration queue was sampled from the earliest chronological portion of Domain C and excluded from the test portion. Transport maps were estimated from this queue using marginal score distributions and acquisition metadata. This experiment tested whether transport calibration can adjust score scale without labels from the new hospital. It is a stricter setting than the internal test, where all domains appear during training.

Ablations examined the proposed model without the ordinal head, without adversarial residual learning, without transport calibration, without curvature penalty, without soft labels, and without draft-impression features. The image-only variant removed all language-derived draft features while retaining acquisition metadata. Another ablation randomized domain labels during training, which tested whether the transport term relied on meaningful domain structure. A final ablation replaced optimal transport with domain-specific temperature scaling, which calibrates probability sharpness but does not align score geometry.

Error review was conducted on 360 test cases sampled from model disagreements. The sample included 120 cases where the proposed model corrected a severe or moderate under-triage error from the softmax baseline, 120 cases where it introduced a new under-triage error, and 120 cases where it produced over-escalation. The review examined image quality, finding type, draft-impression content, portable status, and whether the label was clinically ambiguous. The review was not used to modify the model. It served to interpret statistical results and identify failure modes.

5. Results

The flat softmax baseline achieved 67.8% accuracy, 0.612 macro ordinal agreement, 0.641 quadratic-weighted agreement, and 0.74 mean absolute ordinal error. Class-weighted softmax improved urgent-review sensitivity but reduced specificity, yielding 65.9% accuracy and 0.626 macro ordinal agreement. Focal loss produced 66.4% accuracy and 0.634 macro ordinal agreement. The ordinary ordinal model improved macro ordinal agreement to 0.657 and reduced mean absolute ordinal error to 0.66. Adding adversarial residual learning raised macro ordinal agreement to 0.668 and reduced domain calibration error from 0.102 to 0.081. The full transport-calibrated ordinal model achieved 69.3% accuracy, 0.684 macro ordinal agreement, 0.711 quadratic-weighted agreement, and 0.59 mean absolute ordinal error.

The proposed model did not have the highest ordinary accuracy by a large margin. Its accuracy improvement over the flat softmax baseline was 1.5 percentage points, with a 95% confidence interval from 0.6 to 2.3 points. The larger gains appeared in ordinal and clinical-error metrics. Macro ordinal agreement improved by 0.072, with a confidence interval from 0.058 to 0.086. Mean absolute ordinal error decreased by 0.15. Severe under-triage fell from 6.9% to 3.8% of all true Level 3 cases. Moderate under-triage fell from 18.6% to 13.4%. Severe over-triage rose slightly from 2.7% to 3.1%, but the difference was not statistically reliable. Routine over-escalation changed from 14.9% to 15.7%, remaining below the validation constraint.

Urgent-review sensitivity increased from 81.2% for the flat softmax baseline to 88.5% for the proposed model on the internal test set. Urgent-review specificity decreased from 91.6% to 90.4%. The sensitivity gain was therefore not obtained by a large specificity sacrifice. The class-weighted softmax baseline reached 89.1% urgent sensitivity but specificity dropped to 85.7%, causing many more routine cases to be escalated. The proposed model offered a more balanced shift. At the validation-selected urgent threshold, the positive predictive value for urgent review was 52.8% for the softmax baseline, 43.6% for class-weighted softmax, and 51.2% for the proposed model. The proposed model therefore increased sensitivity while preserving most precision.

For the any-follow-up threshold, the proposed model showed a smaller improvement. Sensitivity rose from 72.5% to 75.1%, and specificity changed from 78.6% to 78.1%. For the expedited-or-higher threshold, sensitivity rose from 69.4% to 75.8%, and specificity changed from 82.2% to 81.3%. These threshold-level results show where ordinal modeling helped. The largest improvement was not at the lowest threshold, where many mild cases are ambiguous, but at the separation between routine and operationally important follow-up [14]. That separation is central to workflow triage because it determines whether a case moves into a higher-priority queue.

Domain calibration error was 0.119 for flat softmax, 0.137 for class-weighted softmax, 0.112 for focal loss, 0.097 for the ordinary ordinal model, 0.081 for the adversarial ordinal model, and 0.063 for the proposed transport-calibrated model. Cross-domain calibration spread decreased from 0.074 for flat softmax

Table 1 Dataset composition across hospital-derived domains.

Metric	Domain A	Domain B	Domain C
Encounters	13,204	12,118	13,290
Portable imaging (%)	39.6	21.8	54.9
Level 3 prevalence (%)	13.7	8.9	10.6
Level 2 prevalence (%)	17.2	15.4	22.4
Main workflow context	Emergency	Outpatient	Portable acute care

Table 2 Four-level ordinal triage definition.

Level	Recommendation	Typical findings	Workflow implication
0	No follow-up	Stable or normal study	Standard reporting
1	Routine follow-up	Mild chronic findings	Outpatient reassessment
2	Expedited follow-up	New opacity or interval concern	Prompt clinician review
3	Urgent review	Pneumothorax or severe effusion	Immediate escalation

to 0.031 for the proposed model. Domain C had the largest calibration error for all methods, likely because it had the highest portable-image rate. The proposed model reduced Domain C calibration error from 0.151 to 0.079 [15]. This reduction matters because a model with similar average accuracy can still be risky if one hospital has systematically inflated or deflated urgency probabilities.

The decision-curve analysis favored the proposed model across most urgent-review thresholds. The area under the net-benefit curve between threshold probabilities 0.05 and 0.50 was 0.184 for flat softmax, 0.176 for class-weighted softmax, 0.191 for focal loss, 0.203 for the ordinary ordinal model, 0.211 for the adversarial ordinal model, and 0.226 for the full model. The advantage was largest between threshold probabilities 0.18 and 0.36, which corresponds to plausible settings where missing urgent cases is costly but over-escalation cannot be unlimited [16]. At very low threshold probabilities, all methods escalated many cases and the difference narrowed. At very high thresholds, sensitivity fell for every model.

The ablation study indicated that no single component explained the full gain. Removing ordinal structure and using a flat head with transport calibration reduced macro ordinal agreement from 0.684 to 0.649. Removing transport calibration reduced it to 0.668 and increased domain calibration error from 0.063 to 0.081 [17]. Removing adversarial residual learning reduced macro ordinal agreement to 0.673 and increased calibration spread. Removing the curvature penalty kept internal performance similar but made held-out-domain adaptation less stable. Removing soft labels reduced performance mainly for Level 1 and Level 2, where label ambiguity was common. The image-only variant reached 0.641 macro ordinal agreement, showing that draft-impression features contributed but did not fully determine performance.

The randomized-domain ablation was especially informative. When domain labels were randomly permuted during training, the transport penalty no longer represented real institutional shift. Macro ordinal agreement fell to 0.661, and domain calibration error rose to 0.093. This result suggests that the transport component was not merely acting as generic regularization. It depended on meaningful grouping by hospital-derived domain. Replacing optimal transport with temperature scaling reduced calibration error relative to no calibration, but not as much as transport. Temperature scaling reached domain calibration error of 0.076, while transport reached 0.063. Temperature scaling adjusted probability sharpness but did not correct threshold-specific score displacement as effectively.

Performance by label level showed a predictable pattern. Level 0 was easiest, with 84.7% class sensitivity for the proposed model. Level 3 was also relatively strong, with 88.5% sensitivity under the urgent threshold and 74.2% exact class sensitivity. Level 1 and Level 2 were harder. Exact sensitivity was 58.6% for Level 1 and 61.9% for Level 2. Many errors between Level 1 and Level 2 reflected genuine ambiguity in report-derived recommendations. The proposed model reduced large-distance mistakes more than adjacent mistakes. The proportion of errors with absolute ordinal distance two or greater fell from 12.8% for softmax to 8.1% for the proposed model.

Finding-level analysis showed the largest improvements for pneumothorax, new focal opacity, and moderate-to-large pleural effusion. For pneumothorax, severe under-triage decreased from 5.4% to 2.1%. For moderate-to-large pleural effusion, it decreased from 8.7% to 4.6%. For new focal opacity, moderate under-triage decreased from 22.3% to 15.2%. For mild vascular congestion, improvement was limited, with Level 1 versus Level 2 confusion remaining common. For support-device-only cases, the proposed model sometimes over-escalated when a tube or line was visually prominent but not clinically urgent. These cases accounted for 18.5% of new severe over-triage errors introduced by the proposed model.

The image-only model had lower performance but similar

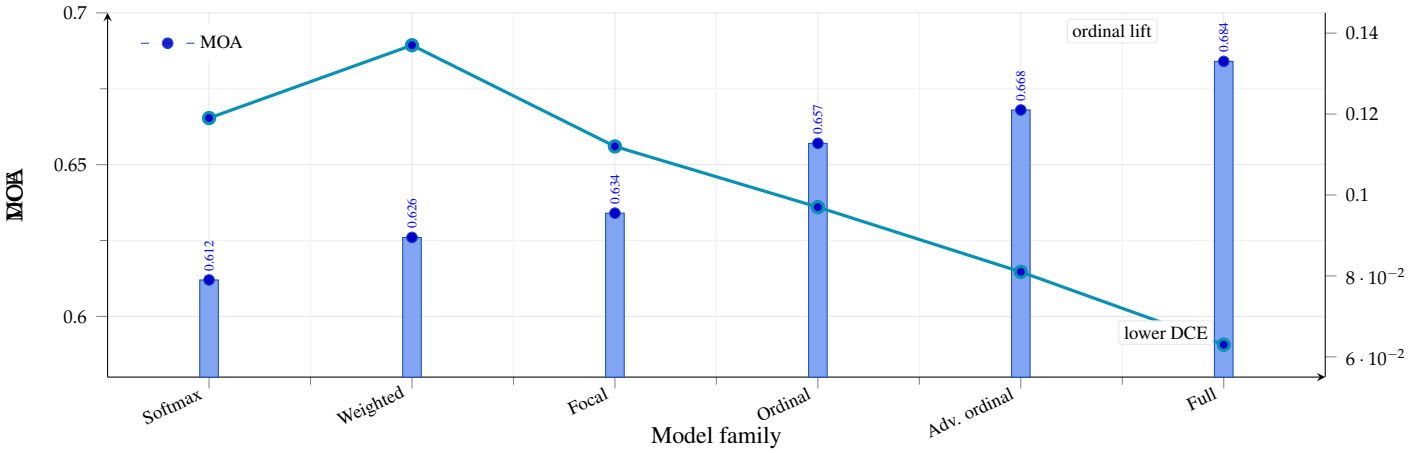

Figure 5 Internal model comparison showing that macro ordinal agreement rises monotonically with structured ordinal and transport components, while domain calibration error falls most strongly in the full model.

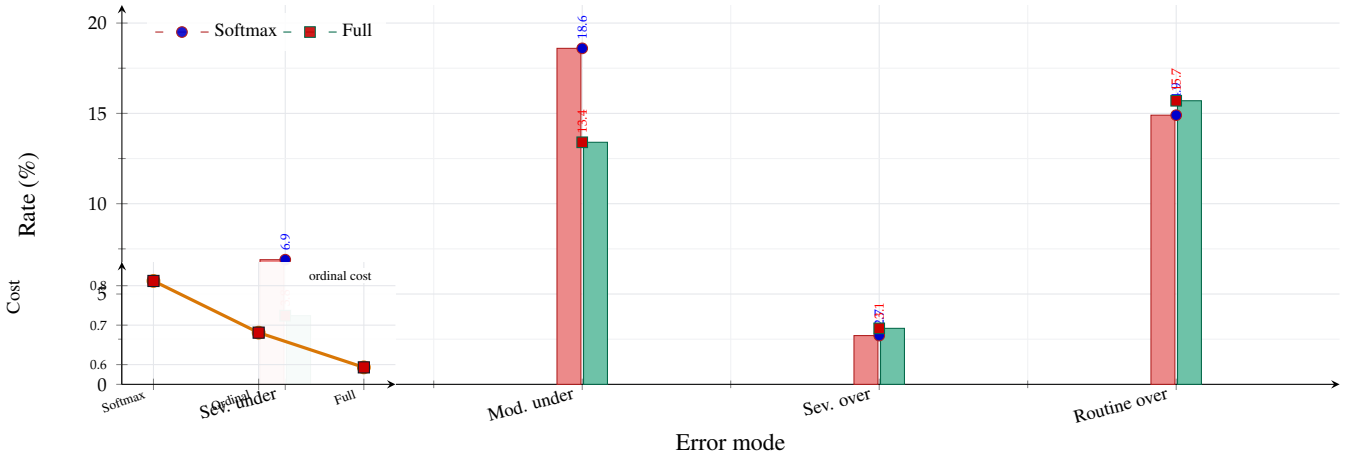

Figure 6 Clinical error geometry shifts toward fewer harmful misses: severe and moderate under-triage decrease, while squared ordinal cost falls from the flat baseline to the proposed model.

error direction. It achieved 64.2% accuracy, 0.641 macro ordinal agreement, and 0.092 domain calibration error. Adding draft-impression features raised performance to the full model's 69.3% accuracy and 0.684 macro ordinal agreement. The improvement from draft features was largest for Level 2 cases, where preliminary language captured uncertainty and follow-up suggestions that were difficult to infer from the image alone. However, draft features also introduced report-style risk. Without adversarial residual learning, the image-plus-draft model showed higher domain calibration spread, indicating that language conventions differed by domain. The proposed method reduced but did not eliminate this issue.

Soft-label training improved performance where reviewers disagreed. On the subset of cases with annotation uncertainty, the hard-label ordinal model had mean absolute ordinal error of 0.91, while the soft-label ordinal model had 0.82. On cases with high reviewer agreement, the difference was small at 0.51 versus 0.50. This suggests that soft labels helped mainly by reducing overconfidence on boundary cases. The proposed full model with soft labels had the best calibration in uncertain cases, with expected cumulative calibration error of 0.081 compared with 0.126 for the hard-label softmax baseline.

The held-out-domain simulation was more difficult [18]. Training on Domains A and B and testing on Domain C reduced performance for every method. Flat softmax reached 63.1% accuracy, 0.574 macro ordinal agreement, and 0.158 domain calibration error on held-out Domain C. The ordinary ordinal model reached 64.5% accuracy, 0.603 macro ordinal agreement, and 0.132 calibration error. The proposed model with unlabeled transport adaptation reached 66.2% accuracy, 0.642 macro ordinal agreement, and 0.088 calibration error. Urgent-review sensitivity increased from 81.2% for softmax to 87.6% for the proposed method, while urgent specificity changed from 88.9% to 88.1%. Severe under-triage fell from 8.2% to 4.9%.

The unlabeled calibration queue size affected held-out performance. With 128 unlabeled encounters, domain calibration error was 0.104. With 256 encounters, it was 0.096. With 512 encounters, it was 0.088. With 1,024 encounters, it improved slightly to 0.084. Macro ordinal agreement changed less, ranging from 0.633 at 128 encounters to 0.646 at 1,024.

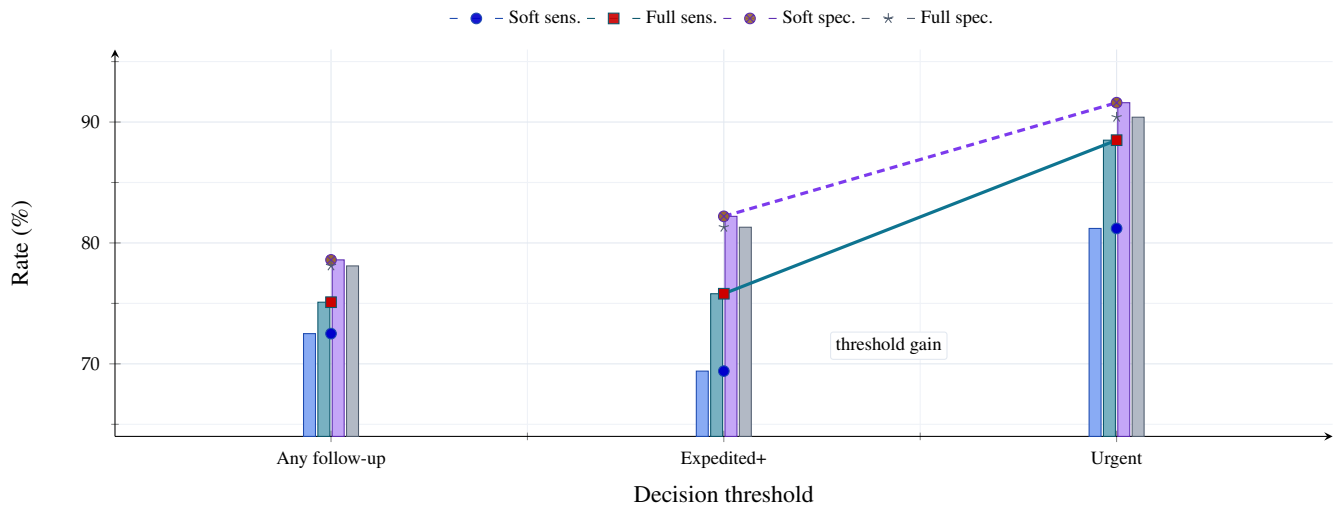

Figure 7 Threshold-level operating points show the main gain at expedited-or-higher and urgent-review boundaries, with sensitivity increasing while specificity remains close to the baseline.

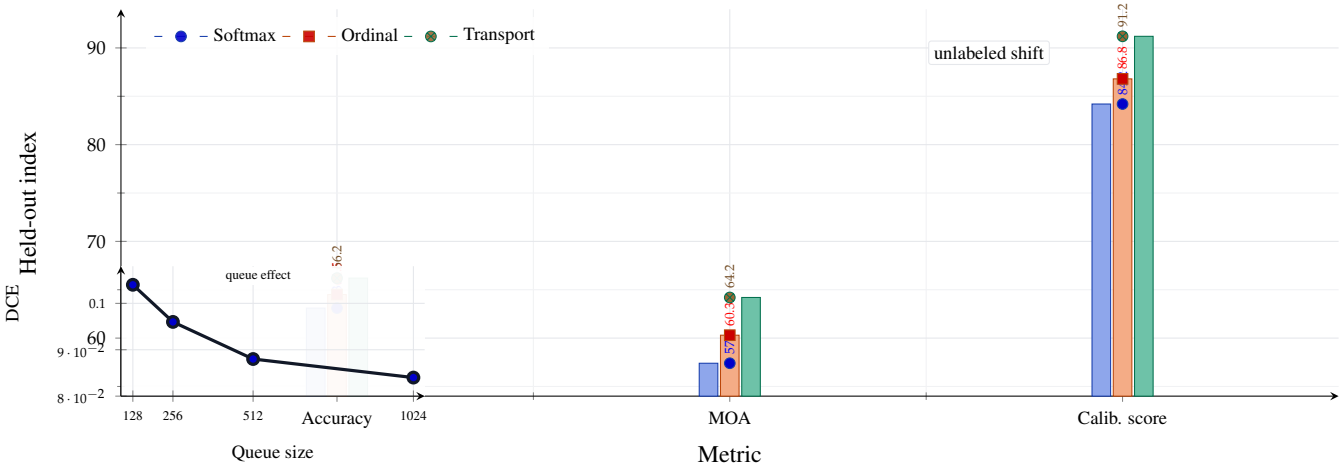

Figure 8 Held-out Domain C evaluation favors unlabeled transport adaptation: accuracy, ordinal agreement, and calibration score improve, while calibration error decreases with larger target queues.

This suggests that a moderate unlabeled sample can estimate score transport sufficiently for calibration, while classification performance depends more on the learned representation. The 512-case queue was chosen as the main setting because it offered most of the calibration benefit without requiring a large unlabeled adaptation period.

6. Clinical Error Geometry

The proposed model changed not only the number of errors but their geometry on the ordinal scale. In the flat softmax model, 34.7% of all mistakes jumped two or more levels. In the proposed model, that proportion was 24.6%. Adjacent mistakes became a larger share of the remaining errors. This shift is clinically meaningful because a Level 3 case predicted as Level 2 is still likely to receive faster attention than a Level 3 case predicted as Level 0 or Level 1. Similarly, a Level 0 case predicted as Level 1 may consume some follow-up resources, but it is less disruptive than direct urgent escalation. The ordinal loss appears to encourage such near-miss behavior.

The severity score distributions showed clearer separation after ordinal training. In the flat softmax model, the implied urgency score was formed from class probabilities and had overlapping distributions for Level 1 and Level 2. The ordinary ordinal model produced a more monotonic arrangement, but Domain C scores were shifted upward relative to Domains A and B [19]. The transport-calibrated model preserved class ordering while reducing domain-specific displacement. The median calibrated score for Level 3 was 2.18, compared with 1.29 for Level 2, 0.47 for Level 1, and -0.61 for Level 0. Interquartile overlap remained substantial between adjacent levels but decreased for non-adjacent levels.

A matrix analysis of threshold errors clarified the source of improved macro ordinal agreement [20]. Let $E \in \mathbb{R}^{4\times4}$ be the empirical confusion matrix with entries E_{ab} for true level a and predicted level b. The expected ordinal cost under squared

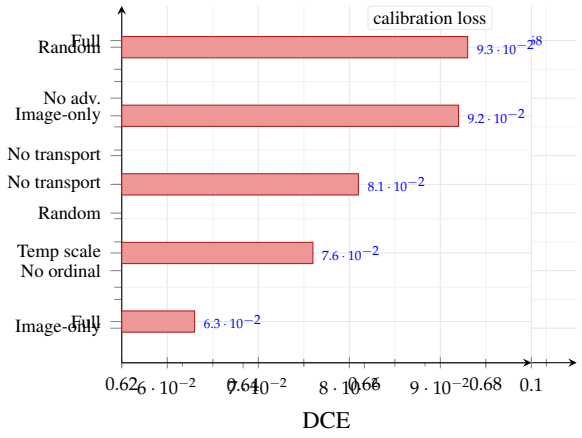

Figure 9 Ablation structure separates performance loss from calibration loss: removing draft features or ordinal structure lowers agreement, while calibration variants mostly affect domain error.

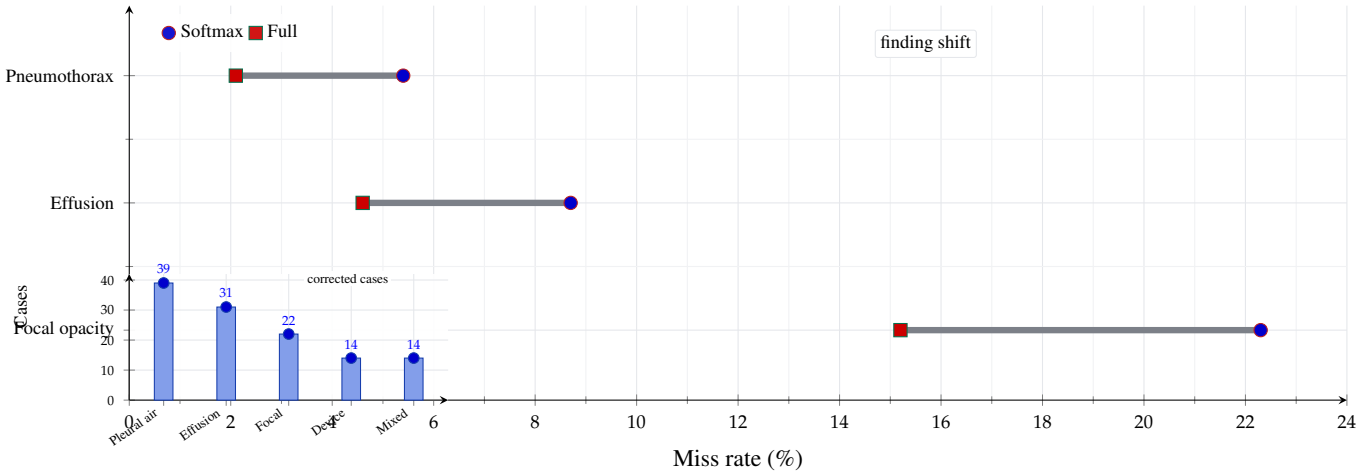

Figure 10 Finding-level analysis shows the largest reductions for pneumothorax, moderate-to-large effusion, and new focal opacity, with corrected cases concentrated in pleural and opacity patterns.

distance is

$$C(E) = \sum_{a=0}^{3} \sum_{b=0}^{3} (a-b)^2 E_{ab}. \tag{11}$$

For flat softmax, $C(E) = 0.812$. For the ordinary ordinal model, $C(E) = 0.681$. For the proposed model, $C(E) = 0.593$. The largest cost reduction came from moving true Level 3 cases away from predicted Level 0 or Level 1 and from moving true Level 2 cases away from predicted Level 0. The cost associated with Level 0 over-escalation increased slightly, but not enough to offset the reduction in under-triage cost. This explains why ordinary accuracy improved modestly while ordinal metrics improved more clearly [21].

To understand which features affected threshold movement, a post hoc linear surrogate was fit to the proposed model's severity scores. The surrogate used structured findings, portable status, image quality, and draft-impression indicators. It explained 63.4% of score variance. Pneumothorax, large pleural effusion, new focal opacity, and malpositioned tube had the largest positive coefficients. Stable cardiomegaly, mild hyperinflation, and chronic scarring had smaller positive coefficients, usually raising cases from Level 0 to Level 1 but not to higher levels. Low image quality had a positive coefficient for expedited follow-up, reflecting model uncertainty and the association between poor acquisition and portable acute imaging. However, in the full model, low quality alone rarely produced Level 3.

The transport map affected score thresholds differently by domain. Domain C raw scores were generally higher because portable images and draft impressions contained more acute-care patterns. Without transport, this caused over-escalation of stable portable cases. The Domain C transport map compressed mid-range scores and preserved high-end scores. In mathematical terms, for raw scores between 0.2 and 1.1, $T_C(r) - r$ averaged -0.24; for raw scores above 1.8, the average shift was -0.04. This selective compression helped reduce routine over-escalation while retaining urgent sensitivity. Domain B showed the opposite pattern: mid-range scores were slightly expanded because outpatient cases with follow-up recommendations were otherwise under-scored.

A case-review sample supported the quantitative pattern.

Table 3 Comparison of baseline and proposed methods.

Method	Accuracy	MOA	Calibration Error	MAOE
Flat softmax	67.8	0.612	0.119	0.74
Weighted softmax	65.9	0.626	0.137	0.72
Focal loss	66.4	0.634	0.112	0.70
Ordinal model	68.2	0.657	0.097	0.66
Adversarial ordinal	68.8	0.668	0.081	0.63
Transport-calibrated ordinal	69.3	0.684	0.063	0.59

Table 4 Clinical error analysis on internal testing.

Metric	Flat Softmax (%)	Proposed Model (%)
Severe under-triage	6.9	3.8
Moderate under-triage	18.6	13.4
Severe over-triage	2.7	3.1
Routine over-escalation	14.9	15.7
Urgent sensitivity	81.2	88.5
Urgent specificity	91.6	90.4

Among 120 cases where the proposed model corrected baseline under-triage, 39 involved visible pleural air or clear pleural-line findings, 31 involved moderate or large effusions, 22 involved new focal opacities with follow-up language in the draft impression, 14 involved support-device concerns, and 14 involved mixed cardiopulmonary findings. In many of these cases, the flat softmax model assigned high probability to Level 1 but failed to cross the expedited or urgent threshold. The ordinal model's cumulative structure made it less likely to stop at routine follow-up when high-threshold evidence was present.

Among 120 cases where the proposed model introduced under-triage, 47 involved subtle or equivocal findings, 29 involved low-quality portable images, 18 involved disagreement between image appearance and draft impression, 15 involved chronic abnormalities with uncertain interval change, and 11 involved rare device complications. These cases show the cost of the model's calibration. It sometimes required stronger evidence to cross higher thresholds, which reduced false urgent calls but missed some ambiguous positives. The problem was most pronounced when the draft impression used cautious language and the image signal was weak. This limitation suggests that uncertainty-aware escalation may need explicit abstention rather than forced classification.

Among 120 over-escalation cases, 38 involved prominent support devices, 27 involved severe chronic cardiomegaly without acute change, 23 involved low-volume portable images that mimicked basilar opacity, 19 involved rotated images, and 13 involved draft-impression phrases that suggested possible follow-up despite benign labels. Transport calibration reduced domain-wide over-escalation but did not fully solve feature-level confounding. Support devices were especially difficult because they can be either incidental or urgent depending on position. A single-view image sometimes lacks enough context to determine that distinction.

The score-threshold analysis revealed that threshold β_2, separating routine from expedited-or-higher follow-up, was the most unstable during training. The first threshold converged early, and the urgent threshold stabilized after class-balanced sampling was introduced. The middle threshold continued to move because Level 1 and Level 2 labels had the most annotation ambiguity. Soft-label training reduced this instability. With hard labels, the standard deviation of β_2 over the last ten epochs was 0.083. With soft labels, it was 0.041. More stable thresholds improved validation calibration and reduced adjacent-level oscillations.

The low-rank factorization showed that six latent directions captured most cumulative-logit variation. The first direction separated no-follow-up cases from all others. The second direction separated urgent cases from the rest and had strong associations with pneumothorax and large effusion. The third direction separated routine chronic findings from acute follow-up cases. The remaining directions captured device concerns, image quality, focal opacity, and diffuse cardiopulmonary burden. This decomposition was not used as definitive explanation, but it supported the view that ordinal triage is not a single pathology detector. It is a structured severity estimate assembled from multiple visual and linguistic factors.

The proposed model's errors were also less sensitive to prevalence changes. In prevalence-stratified simulation, the Level 3 prevalence was artificially varied from 5.0% to 18.0% by resampling the test set. The softmax baseline's urgent positive predictive value ranged from 31.4% to 66.8%, and calibration

Table 5 Threshold-level cumulative performance.

Threshold	Metric	Baseline (%)	Proposed (%)
Any follow-up	Sensitivity	72.5	75.1
Any follow-up	Specificity	78.6	78.1
Expedited-or-higher	Sensitivity	69.4	75.8
Expedited-or-higher	Specificity	82.2	81.3
Urgent review	Sensitivity	81.2	88.5
Urgent review	Specificity	91.6	90.4

Table 6 Held-out Domain C evaluation.

Method	Accuracy	MOA	Calibration Error	Urgent Sensitivity
Flat softmax	63.1	0.574	0.158	81.2
Ordinal model	64.5	0.603	0.132	84.1
Transport-calibrated ordinal	66.2	0.642	0.088	87.6

error increased sharply at the low-prevalence end. The proposed model's positive predictive value ranged from 36.9% to 68.1%, and calibration error remained lower across the range. Sensitivity remained above 85.0% except in the lowest-prevalence sample, where it was 83.7%. This indicates that transport and ordinal calibration improved stability under prevalence movement, though not enough to eliminate prevalence effects.

7. Domain Stress Analysis

The hospital-domain stress tests were designed to separate three kinds of failure: representation failure, threshold failure, and calibration failure. Representation failure occurs when the model cannot identify relevant visual or draft-impression patterns in a new domain. Threshold failure occurs when the model extracts signal but maps it to the wrong triage boundary. Calibration failure occurs when probability estimates are poorly aligned with observed outcomes even if ranking is adequate. The proposed method mainly improved threshold and calibration behavior. It did not fully solve representation failure for unfamiliar image styles or rare findings [22].

In the held-out Domain C experiment, area under the ordinal receiver operating curve for urgent review was 0.914 for the proposed model and 0.902 for the flat softmax model. The ranking difference was modest. Calibration and operating-point behavior differed more. At the selected urgent threshold, the proposed model had higher sensitivity and similar specificity. This indicates that score ordering was already fairly good, but the baseline's probability scale and threshold placement were less appropriate for Domain C. Transport adaptation corrected part of that scale mismatch using unlabeled score distributions.

A threshold-transfer experiment confirmed this interpretation. When thresholds were reselected using labeled Domain C validation data, the softmax baseline's urgent sensitivity improved from 81.2% to 85.4%, narrowing the gap with the proposed model. However, labeled threshold reselection is often unavailable during deployment. The proposed model with unlabeled transport still reached 87.6% urgent sensitivity. This suggests that transport calibration can approximate some benefits of labeled threshold tuning when only unlabeled target-domain data are available. It does not replace full validation but can reduce initial miscalibration.

The transport map was robust to moderate queue contamination. In practice, an unlabeled calibration queue may include cases from mixed scanners, changed protocols, or partial domain labels. To simulate this, the Domain C adaptation queue was contaminated with Domain A cases at rates from 10.0% to 50.0%. Calibration error remained below 0.100 up to 30.0% contamination and rose to 0.117 at 50.0%. Macro ordinal agreement decreased gradually from 0.642 to 0.626. This degradation was smoother than expected because the monotone map was regularized and because Domain A and Domain C shared some acute-care patterns. Stronger contamination or true protocol changes would likely require explicit drift detection.

Temporal drift was simulated by ordering Domain C encounters chronologically and adapting on early cases before testing on later cases. The later portion had a higher rate of portable imaging and a lower rate of routine outpatient follow-up. Without adaptation, domain calibration error was 0.171. With early-queue adaptation, it fell to 0.103. With rolling adaptation using recent unlabeled cases, it fell further to 0.091. The rolling method used no labels but updated score quantiles every 256 cases. It improved calibration but introduced small fluctuations in urgent positive rates. For safety-sensitive deployment, such updates would require monitoring to prevent unintended threshold drift.

Image quality interacted with domain shift. On high-quality held-out images, the proposed model achieved 0.681 macro ordinal agreement. On medium-quality images, it achieved 0.647. On low-quality images, it achieved 0.591. The softmax baseline showed the same ordering but a larger drop, from 0.612

Table 7 Ablation study of proposed components.

Variant	MOA	Calibration Error	Key Effect
Full model	0.684	0.063	Best overall balance
Without ordinal head	0.649	0.089	More large-distance errors
Without transport	0.668	0.081	Poorer calibration
Without adversarial learning	0.673	0.086	Higher domain spread
Without soft labels	0.661	0.078	More Level 1/2 confusion
Image-only variant	0.641	0.092	Lower Level 2 performance

Table 8 Finding-specific severe or moderate under-triage reduction.

Finding Type	Baseline (%)	Proposed (%)
Pneumothorax	5.4	2.1
Moderate-large effusion	8.7	4.6
New focal opacity	22.3	15.2
Mild vascular congestion	16.8	15.1
Support-device cases	11.2	10.4

to 0.498 across high to low quality. Transport calibration helped because low-quality portable images were overrepresented in Domain C and had shifted score distributions. Still, low-quality images remained difficult. The model was especially prone to assigning Level 2 to low-quality cases with uncertain opacity, which may be reasonable in some workflows but can increase follow-up burden.

The draft-impression feature created both benefits and domain risk. In the held-out experiment, removing draft features reduced macro ordinal agreement from 0.642 to 0.611 but also slightly reduced calibration spread. This shows that draft text carries useful clinical signal but also introduces style shift. The adversarial residual component reduced hospital-identifiable information in draft-derived residuals. A domain classifier trained on latent residuals achieved 74.8% accuracy before adversarial training and 46.1% after it, close to the 50.0% chance level for the two-domain held-out training setting. However, domain identity was still partly recoverable from the severity score itself, as expected, because true prevalence differed.

A stress test using altered prevalence but unchanged image distributions showed that transport should not be applied blindly. When Level 3 cases were deliberately oversampled in the unlabeled adaptation queue, marginal transport shifted scores downward and reduced urgent sensitivity. Class-conditional transport would avoid this if labels were available, but the held-out setting used unlabeled data. A prevalence-stabilized variant that matched only the central 70.0% of the score distribution reduced this failure. Under urgent oversampling, ordinary marginal transport reduced sensitivity from 87.6% to 84.2%, while prevalence-stabilized transport maintained 86.5%. This result suggests that target-domain queues should be monitored for case-mix distortion.

The model's calibration was least reliable for Level 2. Across domains, the expected calibration error for the any-follow-up threshold was 0.052, for the expedited-or-higher threshold was 0.081, and for the urgent threshold was 0.057. The middle threshold is clinically and statistically difficult because Level 2 includes heterogeneous cases: possible acute disease, recommended additional imaging, uncertain focal findings, and short-interval follow-up. Some of these cases are visually obvious, while others depend on history or prior imaging. A single-image model cannot fully resolve such ambiguity. The calibration result is therefore not surprising, and it points to the need for prior-study comparison and clinical context.

The domain stress analysis also examined fairness-related metadata in the subset where age and sex were available. The proposed model reduced severe under-triage in all age groups, but absolute rates remained higher in patients above 75 years. This may reflect more chronic abnormalities, more portable imaging, and more complex reports [23]. Severe under-triage in the older group decreased from 8.4% to 5.1%, while in the younger group it decreased from 5.7% to 3.2%. Differences by recorded sex were smaller [24]. The study was not designed as a fairness benchmark, and metadata were incomplete, so these findings should be interpreted as stress checks rather than definitive subgroup conclusions.

The main operational limitation is that transport calibration assumes access to a sufficiently representative unlabeled target stream. A small clinic with few urgent cases may not estimate high-score quantiles well. A hospital undergoing rapid protocol change may have a moving target distribution. A scanner replacement may alter image appearance without immediately changing label prevalence. These conditions can make transport maps stale or misleading. The curvature penalty and monotonicity constraint reduce but do not eliminate this risk. For deployment, transport updates would need monitoring, fallback

Table 9 Effect of unlabeled adaptation queue size.

Queue Size	MOA	Calibration Error	Held-out Setting
128	0.633	0.104	Domain C
256	0.638	0.096	Domain C
512	0.642	0.088	Domain C
1024	0.646	0.084	Domain C

Table 10 Image-quality stress evaluation.

Image Quality	Proposed MOA	Softmax MOA
High quality	0.681	0.612
Medium quality	0.647	0.556
Low quality	0.591	0.498
Portable low-quality subset	0.563	0.471

thresholds, and periodic labeled audits.

The second operational limitation is that follow-up recommendation is partly institutional [25]. Some hospitals escalate certain findings more aggressively than others because of staffing, patient population, or local policy. A model calibrated to a barycentric reference may reduce undesirable technical shift, but it may also smooth away legitimate policy differences. The present study treated the four-level label as a shared target across domains. In real use, hospitals may need local threshold adjustment even if the underlying severity score transfers. The model should therefore be understood as producing a calibrated severity estimate, not as imposing a universal follow-up policy [26].

8. Conclusion

This paper presented a transport-calibrated ordinal approach to chest radiograph follow-up recommendation under hospital shift. The task was formulated as a four-level ordered triage problem rather than a flat classification problem. In the main internal test, the proposed model improved macro ordinal agreement from 0.612 to 0.684, reduced severe under-triage from 6.9% to 3.8%, and reduced domain calibration error from 0.119 to 0.063. In a held-out-domain simulation, it improved macro ordinal agreement from 0.574 to 0.642 and reduced severe under-triage from 8.2% to 4.9%. The results support the view that ordinal structure and domain calibration address different but compatible parts of the triage problem [27].

The most important empirical observation is that ordinary accuracy understated the benefit [28]. Accuracy improved by only 1.5 percentage points over flat softmax, but clinically weighted and ordinal metrics changed more substantially. The model reduced large-distance mistakes, especially cases in which urgent examinations were placed into low-priority categories. It also improved score calibration across hospital-derived domains. These changes are relevant because workflow harm is not evenly distributed across errors. A triage model should be judged by where it places cases on the urgency scale, not only by whether it picks the exact label.

The ablation results showed that the method's components were not interchangeable. The cumulative ordinal head reduced large-distance errors. The residual site-adversarial term reduced domain-specific nuisance structure. The transport layer improved calibration and threshold transfer, especially in the held-out-domain setting. Temperature scaling helped but did not match transport calibration, and randomized domain labels weakened the effect. These findings suggest that the model benefited from structured treatment of hospital shift rather than from generic regularization alone.

The analysis also identified limitations. Level 1 and Level 2 remained difficult to separate, partly because follow-up language and mild imaging findings are inherently ambiguous. The model sometimes under-triaged subtle abnormalities when image quality was low or draft-impression language was cautious. It sometimes over-escalated support-device cases when a device was visually prominent but not clinically urgent. Transport calibration depended on a representative unlabeled target stream and could be affected by case-mix distortion. These limitations matter because a triage system that changes thresholds without monitoring could create new operational risks.

The study should not be read as evidence that automated follow-up recommendation is ready for unsupervised clinical use. The labels were derived from reports and workflow rules, and even the reviewed subset had meaningful ambiguity. The model did not use prior examinations, laboratory data, symptoms, or clinician intent, all of which can change follow-up urgency. The held-out-domain experiment was a controlled simulation rather than a prospective deployment. The results therefore support further evaluation of ordinal transport methods, not direct replacement of clinical triage judgment.

Future work should incorporate longitudinal comparison, richer clinical context, and prospective monitoring. A natural extension is a model that separates imaging severity from local workflow policy, allowing hospitals to tune escalation

thresholds while preserving a shared calibrated score. Another extension is selective prediction, where the model abstains when the middle threshold is uncertain or when the transport map is unstable. The present findings indicate that modeling order and domain movement together can reduce some avoidable triage errors. Whether that improvement translates into safer clinical workflow requires prospective testing with human oversight and institution-specific governance.

Acknowledgments

We would like to thank the reviewers of this research for their helpful comments and suggestions.

References

[1] T. Ding, S. J. Wagner, A. H. Song, R. J. Chen, M. Y. Lu, A. Zhang, A. J. Vaidya, G. Jaume, M. Shaban, A. Kim, D. F. K. Williamson, H. Robertson, B. Chen, C. Almagro-Pérez, P. Doucet, S. Sahai, C. Chen, C. S. Chen, D. Komura, A. Kawabe, M. Ochi, S. Sato, T. Yokose, Y. Miyagi, S. Ishikawa, G. Gerber, T. Peng, L. P. Le, and F. Mahmood, "A multimodal whole-slide foundation model for pathology.," *Nature medicine*, vol. 31, pp. 3749–3761, 11 2025.

[2] B. Sadanandan and V. Behzadan, "Psf-med: Measuring and explaining paraphrase sensitivity in medical vision language models," *arXiv preprint arXiv:2602.21428*, 2026.

[3] D. S. Shrestha, A. M. Straus, R. K. Sah, H. H. Khanal, R. Gautam, B. D. Paudel, T. M. Cornelius, and R. R. Love, "Illness and treatment beliefs in kathmandu valley nepalis under hypertensive care.," *PLOS global public health*, vol. 5, pp. e0004270–e0004270, 6 2025.

[4] A. J. Goodell, S. N. Chu, D. Rouholiman, and L. F. Chu, "Large language model agents can use tools to perform clinical calculations.," *NPJ digital medicine*, vol. 8, pp. 163–163, 3 2025.

[5] M. Lai, I. Y. Chen, J. C. Lai, and J. Ge, "Generative ai in hepatology: Transforming multimodal patient-generated data into actionable insights.," *Hepatology communications*, vol. 9, 7 2025.

[6] N. Shi, Z. Liu, Z. Wan, G. Yang, Y. Shi, P. Wang, and X. Liu, "A vision-language model-based approach for lung cancer diagnosis using lossless 3d ct images: evaluation of gpt-4.1 and gpt-4o for patient-level malignancy assessment.," *Health information science and systems*, vol. 14, pp. 16–16, 12 2025.

[7] C. Liao, F. Wang, S. Li, P. Zhao, W. Wu, Z. Zheng, and L. Shi, "Fostering service-sales ambidexterity: The cross-level influence of self-sacrificial leadership through social exchange mechanisms.," *Acta psychologica*, vol. 254, pp. 104799–104799, 2 2025.

[8] X. Wang, Z. Xiao, and Z. Deng, "Swin attention augmented residual network: a fine-grained pest image recognition method.," *Frontiers in plant science*, vol. 16, pp. 1619551–1619551, 6 2025.

[9] F. Liu, Y. Niu, Q. Zhang, K. Wang, Z. Dong, I. N. Wong, L. Cheng, T. Li, L. Duan, K. Li, G. Li, T. W. Hou, M. Fok, H. Luo, X. Chen, K. Zhang, and Y. Yin, "A foundational architecture for ai agents in healthcare.," *Cell reports. Medicine*, vol. 6, pp. 102374–102374, 9 2025.

[10] Y. Zhang, Y. Xu, P. Chen, S. Wang, Q. Song, L. Yu, and W. Cai, "Knowledge-enhanced medical image classification via descriptive priors from large language models.," *Health information science and systems*, vol. 13, pp. 61–61, 9 2025.

[11] Y. Luo, H. Hooshangnejad, X. Feng, G. Huang, X. Chen, R. Zhang, Q. Chen, W. Ngwa, and K. Ding, "A language vision model approach for automated tumor contouring in radiation oncology.," *Bioengineering (Basel, Switzerland)*, vol. 12, pp. 835–835, 7 2025.

[12] D. Khatri and S. T. Gupta, "Towards comprehensive benchmarking of medical vision language models," *Briefings in Bioinformatics*, vol. 26, pp. i44–45, 12 2025.

[13] J. Shapiro, M. Atlas, S. Baum, F. Pavlotsky, A. Barzilai, R. Gershon, R. Gleicher, and I. Cohen, "One step closer to conversational medical records: Chatgpt parses psoriasis treatments from emrs.," *Journal of clinical medicine*, vol. 14, pp. 7845–7845, 11 2025.

[14] M. Ali, V. Benfante, G. Basirinia, P. Alongi, A. Sperandeo, A. Quattrocchi, A. G. Giannone, D. Cabibi, A. Yezzi, D. D. Raimondo, A. Tuttolomondo, and A. Comelli, "Applications of artificial intelligence, deep learning, and machine learning to support the analysis of microscopic images of cells and tissues.," *Journal of imaging*, vol. 11, pp. 59–59, 2 2025.

[15] X. Sun, J. Liu, L. Wu, X. Chen, X. Ma, F. Teng, T. Zhang, H. Su, X. Fan, J. Li, S. Xu, P. Jin, and H. Jiao, "Ai-powered three-category helicobacter pylori diagnosis via magnetic controlled capsule endoscopy: a multicenter validation of a vision-language model.," *Frontiers in microbiology*, vol. 16, pp. 1687021–1687021, 10 2025.

[16] J. L. Brenner, J. T. Anibal, L. A. Hazen, M. J. Song, H. B. Huth, D. Xu, S. Xu, and B. J. Wood, "Ir-gpt: Ai foundation models to optimize interventional radiology.," *Cardiovascular and interventional radiology*, vol. 48, pp. 585–592, 3 2025.

[17] J. Li, Z. Wang, L. Yu, H. Liu, and H. Song, "Synthetic data-driven approaches for chinese medical abstract sentence classification: Computational study.," *JMIR formative research*, vol. 9, pp. e54803–e54803, 3 2025.

[18] P. Wurster, M. Tharp, K. Ho, and J. Eikenberry, "Incarceration and emergency department visit frequency as predictors for missing day of cataract surgery at county hospital.," *The journal of medicine access*, vol. 9, pp. 27550834251359809–27550834251359809, 7 2025.

[19] H. Hasan, R. Hagerman, D. S. Say, A. P. Nguyen, K. Babata, T. Oyegbile-Chidi, A. Herrera-Guerra, C. Torrents, C. E. Silver, and B. Restrepo, "Parallel paths: A narrative review exploring autism and its co-occurring conditions.," *World journal of clinical pediatrics*, vol. 14, pp. 111641–111641, 12 2025.

[20] G. Liberatore, J. Brenner, J. Franco, and R. Milanaik, "The potential of artificial intelligence to transform medicine.," *Current opinion in pediatrics*, vol. 37, pp. 289–295, 3 2025.

[21] D. Scharp, J. Song, M. H. Palmer, V. Barcelona, and M. Topaz, "Risk factors for emergency department visits or hospitalizations among older adults with urinary incontinence in home health care.," *Journal of gerontological nursing*, vol. 51, pp. 38–47, 4 2025.

[22] L. Dai, H. Xu, and Y. Zhang, "Automated classification of clinical diagnoses in electronic health records using transformer.," *PloS one*, vol. 20, pp. e0329963–e0329963, 9 2025.

[23] L. Zhang, Y. Xing, Y. Pan, Y. Zhao, Z. Wang, Y. Wu, X. Gao, X. Li, Y. Wang, Y. Wang, Y. Guo, Y. Tang, and L. Ma, "Hearing impairment detected by the whisper test is associated with physio-cognitive decline syndrome in community-dwelling older adults.," *The journal of nutrition, health & aging*, vol. 30, pp. 100758–100758, 12 2025.

[24] Q. Liu, R. Deng, C. Cui, T. Yao, Y. Yang, V. Nath, B. Li, Y. Chen, Y. Tang, and Y. Huo, "mtree: Multi-level text-guided representation end-to-end learning for whole slide image analysis.," *IS&T International Symposium on Electronic Imaging*, vol. 37, pp. 183–1–183–7, 2 2025.

[25] G. Perry and M. J. M. J. Tooey, "A decade of does: celebrating the 125th anniversary of mla through an annual meeting conversation with past janet doe lecturers.," *Journal of the Medical Library Association : JMLA*, vol. 113, pp. 116–122, 4 2025.

[26] W. Zhang, Z. Ding, M. Bakaev, O. Razumnikova, M. Kludacz-Alessandri, and J. Wu, "Immersive virtual reality based on head-mounted display in medical education: a systematic review.," *BMC medical education*, vol. 25, pp. 1593–1593, 11 2025.

[27] M. Ahmed, J. Lam, A. Chow, and C.-M. Chow, "A primer on large language models (llms) and chatgpt for cardiovascular healthcare professionals.," *CJC open*, vol. 7, pp. 660–666, 2 2025.

[28] H. Li, J.-F. Fu, and A. Python, "Implementing large language models in health care: Clinician-focused review with interactive guideline.," *Journal of medical Internet research*, vol. 27, pp. e71916–e71916, 7 2025.