Research Article

Beyond Average Error: Calibration, Selective Prediction, and Stress Robustness of A Three-Run Deep Learning Ensemble For Pediatric Bone Age Estimation

Safaa Abass Mohammed*[1], Mohammed Hashim Albashir[2], Amal Yousif Ahmed[3], Suhaila M Elmudather[4]

[1] Department of General Courses, Al-Rayan National College of Nursing, Al-Rayan National Colleges, Al Madinah Al Munawwarah, Saudi Arabia.
[2] General Sciences Department, Al-Rayan National College of Health Sciences, Al-Rayan National Colleges, Al Madinah Al Munawwarah, Saudi Arabia.
[3] Department of Anatomy, Faculty of Medicine, Umm Al-Qura University, Makkah, Saudi Arabia.
[4]Department of Anethesia, Al-Rayan National College of Health Sciences, Al-Rayan National Colleges, PO Box 167, Al Madinah Al Munawarah 41411, Saudi Arabia.
Corresponding author
Safaa Abass Mohammed, sa.youssef@amc.edu.sa

Abstract

Background: Bone-age models are commonly compared by average error, but reliability also depends on repeatability, interval calibration, error ranking, and image robustness.
Purpose: To evaluate the internal reliability of a reproducible three-run ensemble for pediatric bone-age estimation.
Materials and Methods: This retrospective study used 12,611 labeled radiographs from the public RSNA Pediatric Bone Age Challenge. Records were assigned to disjoint training (n=8,827), tuning (n=1,261), calibration (n=1,261), and internal-test (n=1,262) partitions, stratified by sex and age band. Three pretrained ResNet-18 regressors with a sex embedding were trained with identical settings but different random seeds; predictions were averaged. We evaluated error metrics, bootstrap confidence intervals (CIs), split-conformal intervals, disagreement-based selective prediction, and paired image-presentation stress tests.
Results: Individual test MAEs were 8.005, 7.855, and 8.364 months; ensemble MAE was 7.578 months (95% CI, 7.238–7.945), RMSE 9.911 months, and R^2 0.940. The ensemble improved MAE by 0.276 months versus the best member. A fixed 90% split-conformal interval achieved 91.52% coverage with 32.49-month mean width. Disagreement weakly correlated with absolute error (Spearman ρ=0.099). Downsampling, additive noise, and horizontal flipping increased MAE by 3.690, 1.903, and 1.108 months, respectively.
Conclusion: Three-run averaging modestly improved internal accuracy and fixed conformal intervals achieved near-nominal coverage. However, disagreement was inefficient for case-level referral, and stressors exposed important fragility. External testing is required before clinical use.

Keywords: bone age; pediatric radiography; deep learning; ensemble learning; conformal prediction; uncertainty quantification; stress testing.

1. Introduction

Bone age derived from hand–wrist radiographs is used to characterize skeletal maturation in pediatric endocrinology and related clinical settings. Manual interpretation commonly follows the Greulich–Pyle atlas or Tanner–Whitehouse framework and is subject to reference-standard and reader variation.[1] [2] Automated methods have therefore been developed to increase consistency and reduce reporting burden.

The public 2017 RSNA Pediatric Bone Age Challenge accelerated this work by providing a large, common set of hand radiographs with bone-age and sex labels.[3] [4] The challenge stimulated many high-performing systems, but it also made average error on a shared benchmark the dominant measure of progress. Model ensembling was subsequently shown to improve performance, particularly when accurate but diverse models were combined; a four-model ensemble from challenge submissions reduced estimated mean absolute deviation from 4.55 to 3.79 months.[5] Consequently, ensembling itself is not a novel contribution in this domain.

Average error does not fully characterize reliability. A model can maintain an acceptable overall MAE while producing unstable predictions for individual cases or under seemingly modest changes in image presentation. A recent computational stress-testing study demonstrated that an award-winning bone-age model could change predictions substantially under alterations in resolution, orientation, brightness, and contrast, despite acceptable baseline accuracy.[6] Such observations motivate evaluation of robustness, uncertainty, and error distribution rather than reliance on a single leaderboard metric.

Conformal prediction offers a model-agnostic way to transform residuals from a held-out calibration sample into prediction intervals with finite-sample marginal coverage under exchangeability.[7] The guarantee is marginal rather than conditional: it does not imply accurate uncertainty ranking for each patient subgroup, and it need not persist after distribution shift. Likewise, the spread of predictions from independently trained neural networks is often interpreted as a proxy for epistemic uncertainty,[8] but its usefulness for case-level referral should be tested rather than assumed.

The present study does not introduce a new architecture and does not claim external or clinical validation. Instead, it asks four reliability questions within a locked RSNA-only design: whether averaging three independently trained instances improves internal test accuracy; whether residual split-conformal intervals achieve their nominal marginal coverage; whether between-run disagreement identifies high-error cases sufficiently for selective prediction; and whether disagreement and interval behavior remain informative under controlled image-presentation stressors. Reporting was structured with reference to CLAIM 2024 and TRIPOD+AI.[9] [10]

2. Materials and Methods

2.1 Study design and intended use

This was a retrospective methodological model-development and internal-testing study using an openly available, de-identified dataset. The intended scientific use was to evaluate reproducibility, marginal interval calibration, case-level uncertainty ranking, and sensitivity to image-presentation changes. It was not designed to diagnose precocious puberty or growth disorders, and no clinical decision threshold was evaluated because chronological age and clinical outcomes were not available in the RSNA training labels.

The protocol separated model tuning, conformal calibration, and final internal testing. Test labels were not used for model selection, interval fitting, or threshold selection. In accordance with CLAIM terminology, the held-out final partition is described as an **internal test set**, not an external validation cohort.[9]

2.2 Data source and reference standard

The study used the 12,611 labeled hand radiographs released as the training cohort of the 2017 RSNA Pediatric Bone Age Challenge. RSNA reports that the dataset was developed by Stanford University and the University of Colorado and annotated by expert observers; it is available for academic, educational, and other non-commercial use subject to attribution requirements.[3] [4] Each record contained an image identifier, bone age in months, and binary sex. The supplied bone-age label was treated as the reference standard. No attempt was made to reinterpret the radiographs or reconstruct inter-reader variability.

All label records had unique identifiers and could be matched to image files. Because the public file did not include a separate patient identifier, disjointness was enforced at the unique image-identifier level. The study therefore does not make a stronger claim of patient-level disjointness beyond the structure of the released data.

2.3 Data partitions

A deterministic split was generated before model training. The data were divided into training (70%; n=8,827), tuning (10%; n=1,261), calibration (10%; n=1,261), and internal testing (10%; n=1,262). Splitting was stratified jointly by binary sex and four bone-age bands: 0–59, 60–119, 120–179, and ≥180 months. The random seed for partition generation was fixed at 20260827. A locked manifest containing the partition assignment, sex, reference age, and age band was reused throughout all experiments.

The training partition was used for parameter estimation. The tuning partition was used only for early stopping and checkpoint selection. The calibration partition was reserved for conformal quantiles and selective-prediction thresholds. The internal test partition was evaluated only after the three checkpoints were fixed.

2.4 Image preprocessing

Images were opened in grayscale after orientation correction from EXIF metadata, resized to 256×256 pixels with antialiasing, and replicated across three channels. Pixel values were converted to tensors and normalized using the ImageNet channel means (0.485, 0.456, 0.406) and standard deviations (0.229, 0.224, 0.225). No hand segmentation, landmark extraction, cropping, or test-time augmentation was used in the final ensemble evaluation.

The final reported models were trained under empirical risk minimization with the same deterministic resizing and normalization pipeline used for tuning. A separate augmentation experiment was explored during development but was discarded before final testing because it worsened tuning performance; it is not part of the confirmatory results.

2.5 Model architecture

Each ensemble member used an ImageNet-pretrained ResNet-18 encoder.[11] The original classification layer was replaced by an identity mapping. Binary sex was represented through an eight-dimensional learned embedding and concatenated with the 512-dimensional image feature vector. The regression head comprised layer normalization, dropout (0.15), a fully connected layer with 128 units, Gaussian error linear unit activation, a second dropout layer (0.15), and a scalar output in months.

All three members had the same architecture, data partitions, and hyperparameters. Diversity arose only from training stochasticity under seeds 101, 202, and 303. The final point estimate was the unweighted arithmetic mean of the three

predictions. This same-architecture ensemble was intended to quantify sensitivity to random training variation; it should not be interpreted as equivalent to a heterogeneous multi-architecture ensemble.[5]

2.6 Training and checkpoint selection

Models were optimized with AdamW using a learning rate of 3×10^{-4}, weight decay of 1×10^{-4}, batch size 64, and a cosine-annealing learning-rate schedule. The objective was Smooth L1 loss with β=6 months. Training used automatic mixed precision on a Tesla T4 GPU for a maximum of 15 epochs. After every epoch, MAE and RMSE were computed on the tuning partition. The checkpoint with the lowest tuning MAE was retained. Early stopping was configured with a patience of four epochs without improvement.

Random seeds controlled Python, NumPy, and PyTorch operations where supported. The three retained checkpoints had best tuning MAEs of 7.874, 7.914, and 8.068 months for seeds 101, 202, and 303, respectively. The corresponding best epochs were 15, 15, and 14.

2.7 Primary performance analysis

The primary endpoint was MAE in months for the ensemble mean on the internal test set. Secondary measures were RMSE, R^2, mean signed error, residual standard deviation, and the percentages of predictions with absolute error ≤6 and ≤12 months. Each member and the ensemble were evaluated on the same cases.

Case-level nonparametric bootstrap resampling with 5,000 replicates and seed 20260829 was used to estimate 95% percentile CIs for model MAE. Paired bootstrap resampling was used to compare the absolute errors of each member with those of the ensemble. The effect was reported as member MAE minus ensemble MAE; a positive value favored the ensemble.

2.8 Conformal prediction intervals

Symmetric split-conformal intervals were fitted on the calibration partition at nominal coverage levels of 80%, 90%, and 95%. For the fixed method, the nonconformity score was the absolute residual, $|y-\hat{y}|$. The finite-sample quantile used rank $\lceil (n+1)(1-\alpha) \rceil$ among n calibration scores. The interval for a test prediction was $\hat{y}\pm q$.

A secondary disagreement-adaptive analysis divided each calibration residual by $s+\varepsilon$, where s was the standard deviation of the three member predictions and $\varepsilon=0.25$ month. The resulting quantile was multiplied by $s+\varepsilon$ at test time. This analysis tested, rather than assumed, whether between-run disagreement provided an efficient case-specific scale. Empirical coverage and mean and median interval widths were reported. Coverage under transformed images was descriptive because exchangeability with the clean calibration set cannot be assumed.

2.9 Selective-prediction analysis

Between-run standard deviation was evaluated as a deferral score; lower values indicated greater apparent confidence. Thresholds corresponding to nominal acceptance fractions of 25%, 50%, 75%, and 100% were determined only from the calibration distribution. These fixed thresholds were then applied to the internal test set and each stress condition. Realized coverage, accepted-case MAE, and the proportions of accepted cases with errors greater than 6 and 12 months were recorded.

The association between disagreement and absolute error was summarized by Spearman rank correlation. Selective-prediction results were interpreted descriptively and were not treated as evidence of clinical safety.

2.10 Image-presentation stress tests

Stress tests were applied to every internal test image before deterministic resizing and normalization. The prespecified moderate conditions were rotation by 7°, brightness multiplication by 1.16 and 0.84, contrast multiplication by 1.20 and 0.80, Gaussian blur with radius 0.75, deterministic additive sinusoidal intensity noise with peak amplitude 0.02 of the 8-bit range, downsampling to 128×128 pixels followed by upsampling to the original size, and upper-left corner occlusion covering 8% of the smaller image dimension. Horizontal flipping was added as an adverse orientation test.

For each condition, the same three checkpoints generated an ensemble prediction. We calculated transformed-image MAE, absolute prediction shift relative to the clean image, between-run disagreement, adaptive interval behavior, and selective-prediction performance. The paired change in case-level absolute error relative to the clean image was summarized with 5,000-replicate bootstrap 95% CIs. These transformations were presentation stressors, not validated simulations of specific acquisition faults or pathology.

2.11 Subgroup analysis

Clean internal-test performance was summarized by sex and the four prespecified age bands. We reported sample size, MAE, RMSE, signed error, fixed 90% conformal coverage, and adaptive 90% coverage. The analyses were descriptive; no claim of demographic fairness was made because race, ethnicity, institution, and other relevant variables were unavailable.

2.12 Software and reproducibility

Training and analysis were implemented in Python 3.13 with PyTorch 2.11.0 and torchvision 0.26.0. Data handling and analysis used pandas, NumPy, Pillow, scikit-learn, and tqdm. The code package includes deterministic split generation, data auditing, training, ensemble evaluation, conformal calibration, stress transformations.[9] [10]

2.13 Ethics

The analysis used only the public, de-identified RSNA challenge dataset and involved no participant contact or locally collected clinical data.

3. Results

3.1 Clean internal-test performance

The internal test set contained 1,262 images. Individual-run MAE ranged from 7.855 to 8.364 months. The ensemble mean achieved an MAE of 7.578 months (95% CI, 7.238–7.945), RMSE of 9.911 months, and R^2 of 0.940. Its mean signed error was 0.310 month, and 48.49% and 80.59% of predictions were within 6 and 12 months of the reference standard, respectively. The corresponding performance comparison is shown in Table 1 and Fig. 1.

Table 1. Clean internal-test performance

Model	MAE, months (95% CI)	RMSE, months	R^2	Mean signed error, months	Within 6 months, %	Within 12 months, %
Seed 101	8.005 (7.631–8.390)	10.533	0.932	0.382	47.31	78.92
Seed 202	7.855 (7.470–8.231)	10.475	0.933	−0.050	47.54	78.13
Seed 303	8.364 (7.986–8.744)	10.785	0.928	0.598	45.56	75.83
Ensemble mean	7.578 (7.238–7.945)	9.911	0.940	0.310	48.49	80.59

The ensemble reduced MAE by 0.427 month relative to seed 101 (95% CI, 0.239–0.614), 0.276 month relative to seed 202 (95% CI, 0.088–0.473), and 0.786 month relative to seed 303 (95% CI, 0.594–0.972). Thus, the ensemble outperformed every member on the common test cases, although the gain over the best member was modest.

3.2 Conformal interval calibration

The calibration results are summarized in Table 2 and Fig. 5. Fixed split-conformal intervals achieved empirical coverages of 81.78%, 91.52%, and 95.09% at nominal levels of 80%, 90%, and 95%, respectively. The corresponding mean widths were 24.51, 32.49, and 39.64 months. Disagreement-adaptive intervals were wider at all levels and undercovered at the 80% and 90% targets.

Table 2. Conformal prediction-interval performance on clean internal-test images

Method	Nominal coverage	Empirical coverage	Mean interval width, months
Fixed split-conformal	80%	81.78%	24.51
Disagreement-adaptive	80%	78.84%	30.23
Fixed split-conformal	90%	91.52%	32.49
Disagreement-adaptive	90%	88.91%	47.61
Fixed split-conformal	95%	95.09%	39.64
Disagreement-adaptive	95%	95.17%	69.93

At nominal 90% coverage, the adaptive interval was approximately 15.11 months wider on average than the fixed interval while achieving lower empirical coverage. Between-run disagreement therefore did not provide an efficient heteroscedastic scaling factor in this setting.

3.3 Disagreement and selective prediction

Between-run standard deviation correlated weakly with absolute ensemble error (Spearman ρ=0.099); the correlation for the prediction range was 0.100. Error increased in the highest uncertainty deciles, but the relationship was not monotonic across the full distribution. The corresponding association is depicted in Fig. 4.

Calibration-derived thresholds yielded a realized acceptance fraction of 52.06% for the nominal 50% threshold, with accepted-case MAE of 7.104 months. At the nominal 25% threshold, 26.94% of cases were accepted and MAE was 6.854 months. The reduction was limited relative to the large number of deferred cases and did not establish a clinically actionable operating point. The resulting risk–coverage profile is shown in Fig. 3.

3.4 Image-presentation stress robustness

Table 3 and Fig. 2 summarize these stress-test results. The ensemble was most sensitive to resolution reduction, deterministic additive noise, and horizontal flipping. Downsampling to 128 pixels increased MAE from 7.578 to 11.269 months, a paired increase of 3.690 months (95% CI, 3.186–4.206). The mean absolute prediction shift was 8.427 months, and the 95th percentile shift was 20.058 months. Additive noise increased MAE by 1.903 months (95% CI, 1.435–2.420), while horizontal flipping increased it by 1.108 months (95% CI, 0.814–1.405).

Table 3. Paired change in MAE under image-presentation stress

Stress condition	MAE under stress, months	Change versus clean, months (95% CI)	Mean absolute prediction shift, months
Downsample to 128 pixels	11.269	3.690 (3.186–4.206)	8.427
Deterministic additive noise, amplitude 0.02	9.482	1.903 (1.435–2.420)	4.654
Horizontal flip	8.687	1.108 (0.814–1.405)	4.748
Contrast increase, 20%	7.668	0.090 (0.012–0.169)	1.160
Brightness decrease, 16%	7.663	0.085 (0.015–0.152)	0.888
Contrast decrease, 20%	7.658	0.080 (0.011–0.148)	0.885
Gaussian blur, radius 0.75	7.593	0.014 (−0.004–0.032)	0.252
Upper-left corner occlusion, 8%	7.592	0.013 (0.001–0.025)	0.113
Brightness increase, 16%	7.576	−0.002 (−0.063–0.057)	0.812
Rotation, 7°	7.462	−0.117 (−0.303–0.066)	2.744

Disagreement increased under downsampling and additive noise, but the adaptive intervals became very wide. The mean widths of nominal 90% adaptive intervals were 92.59 months after downsampling and 94.46 months after additive noise. Increased disagreement therefore indicated distributional sensitivity but did not provide a practically efficient interval.

3.5 Subgroup results

Male and female MAEs were 7.434 and 7.749 months, respectively. The 120–179-month group had the lowest MAE (7.133 months), whereas the ≥180-month group had the highest MAE (9.052 months). Mean signed error suggested overestimation at 0–59 months (4.550 months) and underestimation at ≥180 months (−5.711 months), consistent with regression toward the center of the age distribution. Detailed subgroup results are provided in Table 4 and Fig. 6.

Table 4. Clean internal-test performance by sex and age band

Group	n	MAE, months	Mean signed error, months	Fixed 90% coverage
Female	578	7.749	1.135	90.31%
Male	684	7.434	−0.388	92.54%
0–59 months	79	8.279	4.550	92.41%
60–119 months	350	7.878	3.046	89.43%
120–179 months	723	7.133	−0.563	92.67%
≥180 months	110	9.052	−5.711	90.00%

4. Discussion

4.1 Principal findings

This internal reliability study yielded four main findings. First, an unweighted average of three independently trained instances improved MAE relative to each instance, including a 0.276-month improvement over the best member. Second, fixed residual split-conformal intervals achieved marginal coverage close to their nominal levels. Third, between-run disagreement was a weak case-level error-ranking signal and produced adaptive intervals that were wider and not better calibrated than fixed intervals. Fourth, downsampling, additive noise, and horizontal flipping exposed substantial prediction fragility despite acceptable clean-set performance.

The study therefore provides both positive and negative evidence. Ensembling reduced random-training sensitivity, and fixed conformal calibration added an interpretable interval around the point estimate. However, the commonly intuitive assumption that greater ensemble spread necessarily identifies the cases with greatest error was not supported strongly enough for clinical referral.

4.2 Relation to prior work

Pan et al. demonstrated that heterogeneous ensembles from the RSNA challenge improved bone-age accuracy, especially when member errors were less correlated.[5] The present ensemble is deliberately narrower: all members share the same architecture and differ only by training seed. Its smaller gain over the best member is therefore unsurprising. The result is best interpreted as evidence that averaging can stabilize a reproducible baseline, not as a new ensembling method or a state-of-the-art accuracy claim.

Santomartino et al. showed that a high-performing bone-age system could remain vulnerable to clinically plausible image transformations.[6] Our findings agree most clearly for resolution changes and flipping. Moderate downsampling increased MAE by almost four months, with a 95th-percentile prediction shift exceeding 20 months. The comparatively small effects of moderate brightness, contrast, blur, and corner occlusion emphasize that robustness is transformation- and severity-specific rather than a single model property.

Fixed split-conformal intervals behaved as expected under the clean internal exchangeability setting, whereas disagreement-adaptive intervals were inefficient. Conformal coverage is marginal and contingent on exchangeability; it should not be interpreted as a guarantee for each age band or under transformed images.[7] The observed behavior under stress—greater disagreement accompanied by very wide adaptive intervals—illustrates the distinction between detecting that inputs have changed and estimating a useful patient-specific error bound.

4.3 Implications

A practical implication is that bone-age studies should report more than point-estimate accuracy. At minimum, internal evaluations can include repeated training, case-bootstrap uncertainty around MAE, a dedicated calibration partition, interval coverage and width, subgroup error distributions, and paired sensitivity analyses under relevant image-presentation changes. CLAIM 2024 specifically emphasizes complete descriptions of partitions, model training, ensembling, uncertainty, and robustness.[9]

The results do not justify automated triage. Even on clean images, fewer than half of predictions were within six months of the reference standard. A nominal 90% conformal interval had a mean width of 32.49 months, which may be too broad for many intended uses. Selective prediction based on between-run disagreement required deferring nearly half the cases to reduce MAE by less than half a month. Such findings argue for keeping human interpretation central and for performing external testing before any clinical use.

4.4 Strengths and limitations

Strengths include a fully separated tuning, calibration, and testing design; three independent training runs; case-level paired comparisons; finite-sample conformal quantiles; fixed calibration-derived deferral thresholds; subgroup summaries; and deterministic stress transformations applied to every internal test image. The complete case-level predictions and code pathway permit audit and reanalysis.

The principal limitation is the absence of external testing. All partitions originated from the same public RSNA cohort, so the study cannot establish transportability to another institution, population, device, or acquisition protocol. The public labels provide limited demographic information, and chronological age and clinical outcomes are unavailable; consequently, the study cannot evaluate advanced or delayed maturation, precocious puberty, growth delay, or clinical decision utility. Disjointness was enforced at image-ID level because a separate patient identifier was unavailable.

The ensemble used three instances of one architecture and therefore had limited representational diversity. Its accuracy was below that of the strongest heterogeneous challenge ensembles, and no comparison was made with contemporary transformer or multi-region models. The adaptive conformal analysis used only between-run standard deviation and may not generalize to other uncertainty scores. The stress suite was computational and did not reproduce scanner physics;

the deterministic additive noise should not be equated with a validated acquisition-noise model. Finally, subgroup estimates for the youngest and oldest age bands were based on 79 and 110 cases, respectively, and should be interpreted cautiously.

5. Conclusion

Averaging three independently trained ResNet-18 bone-age regressors improved internal test accuracy and reduced sensitivity to a single training run. Fixed residual split-conformal intervals achieved useful marginal calibration, but between-run disagreement was only weakly associated with case-level error and produced inefficient adaptive intervals. Downsampling, additive noise, and horizontal flipping revealed substantial fragility not captured by clean-set MAE alone. These findings support a multidimensional reliability assessment for bone-age AI and argue against clinical claims without external testing, stronger uncertainty measures, and institution-specific evaluation.

Declarations

Ethics approval and consent to participate

This retrospective analysis used only the publicly available, de-identified RSNA Pediatric Bone Age Challenge dataset. No local recruitment, local medical records, reidentification, intervention, or participant contact occurred. Consent to participate was not applicable.

Consent for publication

Not applicable.

Data availability

The RSNA Pediatric Bone Age Challenge dataset is available from RSNA for academic, educational, and other non-commercial use subject to its terms and attribution requirements.[3] The present analysis used the released training images and annotations. A blinded code package and locked split manifest are supplied for peer review; the authors will deposit a public versioned release on acceptance.

Declaration of generative AI and AI-assisted technologies in the manuscript preparation process

During the preparation of this work, the authors used AI to review language editing, and code. After using this tool, the authors reviewed and edited the content as needed and took full responsibility for the content of the published article.

Acknowledgements

The authors acknowledge the Radiological Society of North America and the institutions and expert annotators who developed the Pediatric Bone Age Challenge dataset.

Funding

This research did not receive any grant from funding.

Declaration of competing interests

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Figure legends

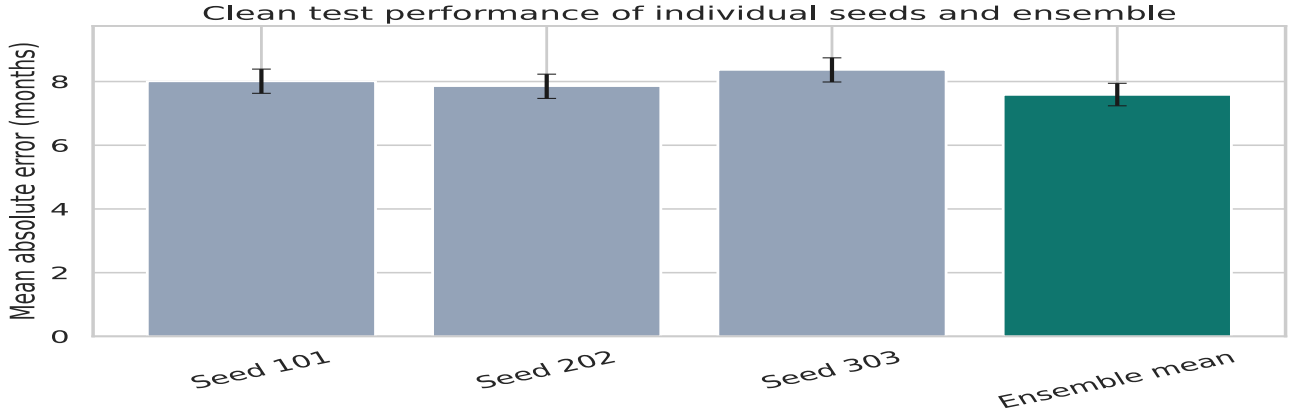

Figure 1. Clean internal-test performance of individual runs and ensemble. Error bars represent case-bootstrap 95% confidence intervals for mean absolute error (MAE).

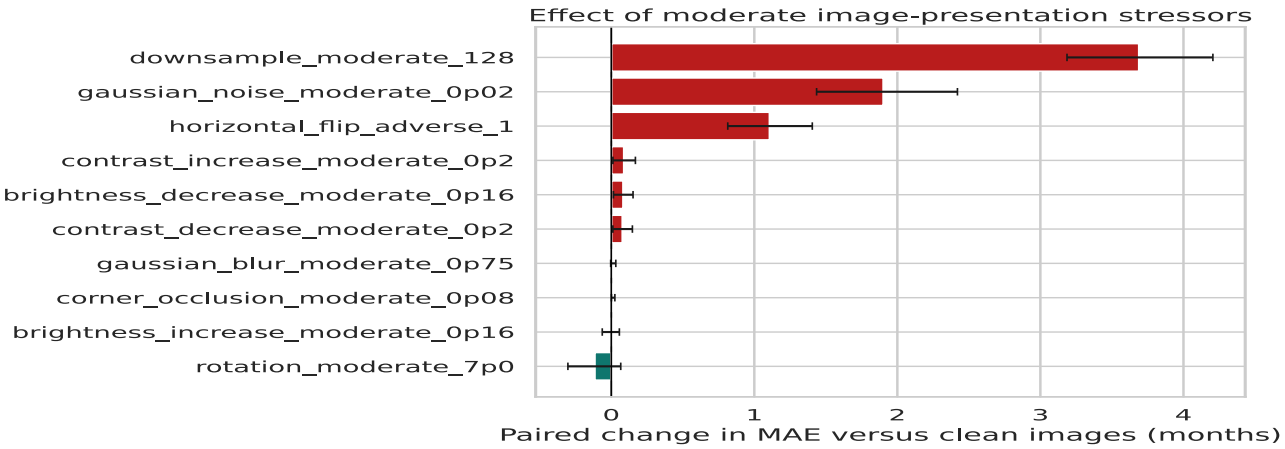

Figure 2. Paired changes in MAE under image-presentation stress. Positive values indicate increased error relative to clean images; bars show paired-bootstrap 95% confidence intervals.

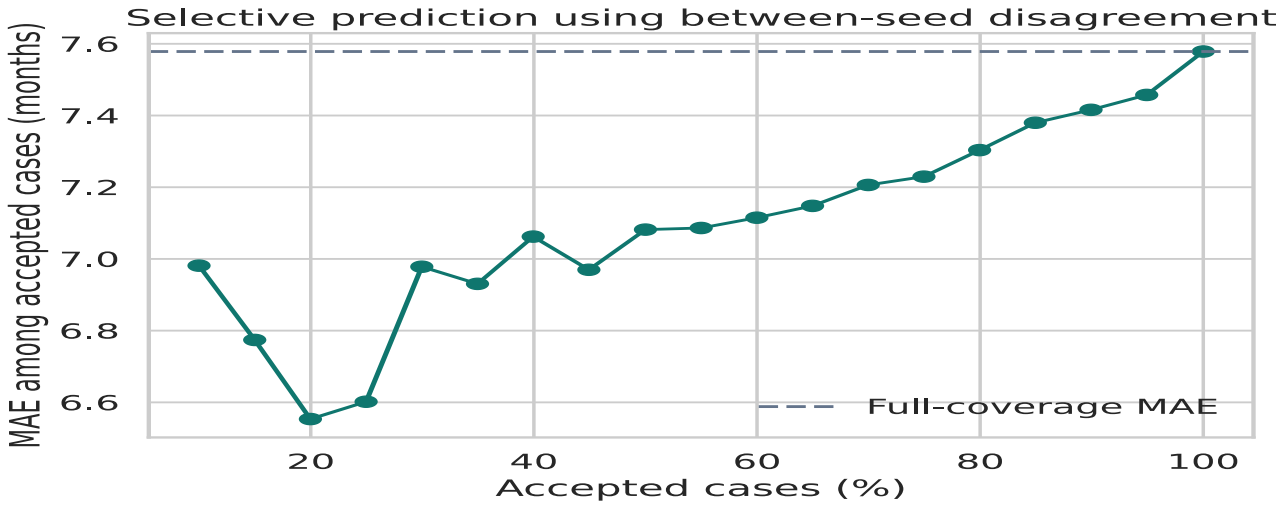

Figure 3. Risk–coverage curve for selective prediction using between-run disagreement. The dashed reference line denotes ensemble MAE at full coverage.

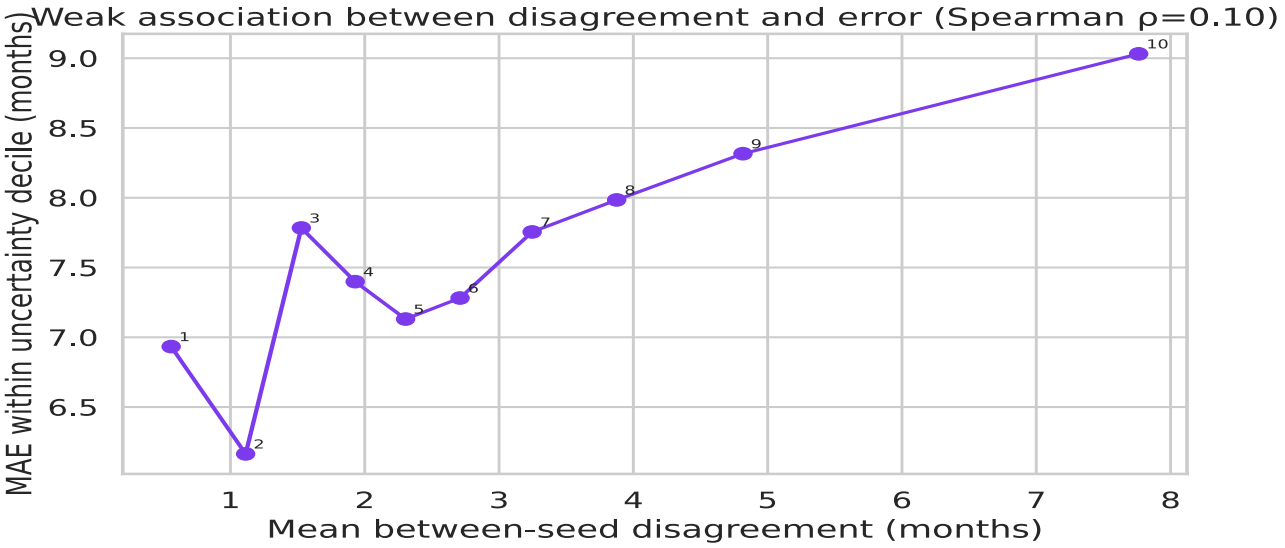

Figure 4. Association between between-run disagreement and absolute error. Points represent uncertainty deciles; the nonmonotonic pattern reflects the weak Spearman correlation.

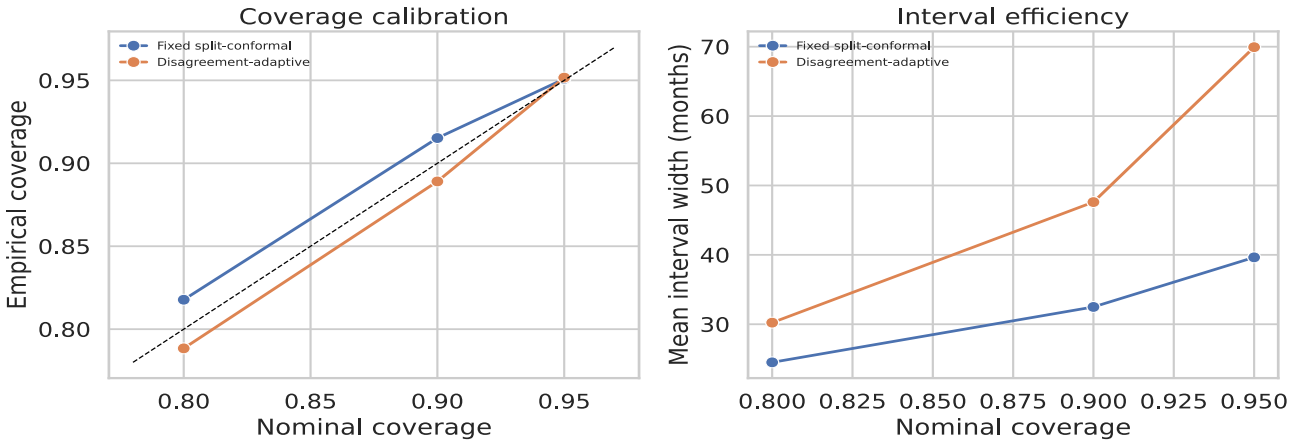

Figure 5. Coverage and efficiency of fixed and disagreement-adaptive split-conformal intervals across nominal coverage levels.

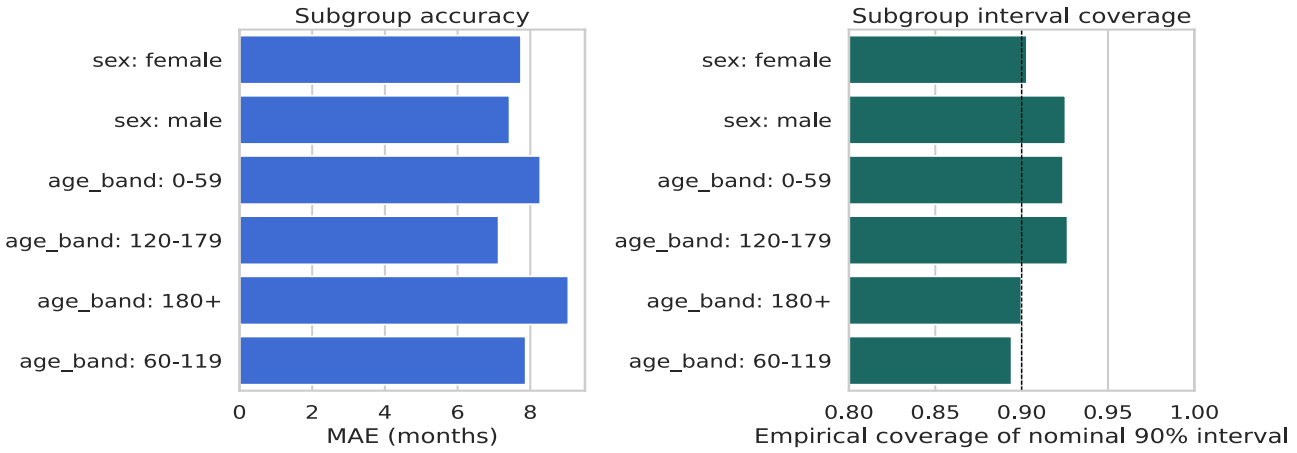

Figure 6. Clean internal-test MAE and fixed 90% split-conformal empirical coverage by sex and bone-age band. Subgroup analyses are descriptive.

References

1. Greulich WW, Pyle SI.Radiographic Atlas of Skeletal Development of the Hand and Wrist. Stanford University Press; 1959.
2. Tanner JM, Healy MJR, Goldstein H, Cameron N. Assessment of Skeletal Maturity and Prediction of Adult Height: TW3 Method. Saunders; 2001.
3. Radiological Society of North America. RSNA Pediatric Bone Age Challenge (2017). Accessed August 2026.https://www.rsna.org/artificial-intelligence/ai-image-challenge/rsna-pediatric-bone-age-challenge-2017 (https://www.rsna.org/artificial-intelligence/ai-image-challenge/rsna-pediatric-bone-age-challenge-2017)
4. Halabi SS, Prevedello LM, Kalpathy-Cramer J, et al. The RSNA Pediatric Bone Age Machine Learning Challenge. Radiology. 2019;290(2):498–503. doi:10.1148/radiol.2018180736.https://pubmed.ncbi.nlm.nih.gov/30480490/ (https://pubmed.ncbi.nlm.nih.gov/30480490/)
5. Pan I, Thodberg HH, Halabi SS, Kalpathy-Cramer J, Larson DB. Improving automated pediatric bone age estimation using ensembles of models from the 2017 RSNA Machine Learning Challenge. Radiology: Artificial Intelligence. 2019;1(6):e190053. doi:10.1148/ryai.2019190053.https://pmc.ncbi.nlm.nih.gov/articles/PMC6884060/ (https://pmc.ncbi.nlm.nih.gov/articles/PMC6884060/)
6. Santomartino SM, Putman K, Beheshtian E, Parekh VS, Yi PH. Evaluating the robustness of a deep learning bone age algorithm to clinical image variation using computational stress testing. Radiology: Artificial Intelligence.

2024;6(3):e230240. doi:10.1148/ryai.230240.https://pmc.ncbi.nlm.nih.gov/articles/PMC11140516/ (https://pmc.ncbi.nlm.nih.gov/articles/PMC11140516/)
7. Lu C, Lemay A, Chang K, Höbel K, Kalpathy-Cramer J. Fair conformal predictors for applications in medical imaging. arXiv:2109.04392. Revised 2022.https://arxiv.org/abs/2109.04392 (https://arxiv.org/abs/2109.04392)
8. Lakshminarayanan B, Pritzel A, Blundell C. Simple and scalable predictive uncertainty estimation using deep ensembles. In: Advances in Neural Information Processing Systems 30. 2017.https://proceedings.neurips.cc/paper/2017/hash/9ef2ed4b7fd2c810847ffa5fa85bce38-Abstract.html (https://proceedings.neurips.cc/paper/2017/hash/9ef2ed4b7fd2c810847ffa5fa85bce38-Abstract.html)
9. Tejani AS, Klontzas ME, Gatti AA, et al. Checklist for Artificial Intelligence in Medical Imaging (CLAIM): 2024 update. Radiology: Artificial Intelligence. 2024;6(4):e240300. doi:10.1148/ryai.240300.https://pmc.ncbi.nlm.nih.gov/articles/PMC11304031/ (https://pmc.ncbi.nlm.nih.gov/articles/PMC11304031/)
10. Collins GS, Moons KGM, Dhiman P, et al. TRIPOD+AI statement: updated guidance for reporting clinical prediction models that use regression or machine learning methods. BMJ. 2024;385:e078378. doi:10.1136/bmj-2023-078378.https://www.bmj.com/content/385/bmj-2023-078378 (https://www.bmj.com/content/385/bmj-2023-078378)
11. He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2016:770–778.https://doi.org/10.1109/CVPR.2016.90 (https://doi.org/10.1109/CVPR.2016.90)