On the Behavior of Calibration Error Under Rater Disagreement

Harshit Kumar*, Yen Nhi Truong Vu, Jason Su, and Thomas Paul Matthews

Whiterabbit.ai, Redwood City, CA, USA
*harshit@whiterabbit.ai

Abstract. Medical AI systems are often evaluated on datasets where ground truth labels suffer from inter-rater variability. Calibration is typically computed against majority-vote consensus to assess whether predicted probabilities reflect correctness. However, recent work argues that calibration to human majority is theoretically flawed in inherently subjective tasks, calling into question the utility of calibration error in such settings. In this work, we demonstrate that this flaw is not with calibration itself but arises from label aggregation strategies, such as majority voting, that induce a mismatch between the calibration target and the intended meaning of the model predictions. Specifically, we show that this mismatch breaks the oracle guarantee that a predictor returning the true probabilities should achieve zero calibration error in expectation. We then consider two principled alternatives to eliminate this mismatch: (i) computing calibration directly against individual rater labels, and (ii) retaining consensus labels while analytically aligning predictions with the aggregation rule. We prove that both approaches recover consistent calibration estimates that restore the oracle guarantee, showing that calibration remains meaningful under label stochasticity. Through controlled sensitivity analyses and a real-world breast density case study, we empirically validate the theoretical behavior of the two protocols. Together, this work establishes the first principled framework that demonstrates the oracle guarantee of calibration error under subjective labeling.

Keywords: Evaluation · Calibration · Rater subjectivity

1 Introduction

AI systems are increasingly deployed for medical imaging tasks that directly influence clinical decisions [20]. Unlike many benchmark machine learning settings [26], clinical ground truth is often inherently subjective, with reasonable disagreement among qualified raters [16,4]. During testing, multi-rater annotations are commonly reduced to a single case-level majority-vote label.

Evaluation metrics, including calibration error, are then computed against this binary label. Calibration assesses whether predicted probabilities match observed outcome frequencies, that is, whether a model meaningfully quantifies uncertainty [17,6]. However, recent work argues that calibration against majority vote is theoretically flawed in subjective tasks [3], as even an oracle predictor

may appear miscalibrated under consensus evaluation. This has led to claims that calibration error should *not* be used in inherently subjective settings [3]. Rather than abandoning calibration error, we examine whether its theoretical validity under inter-rater variability can be restored.

We address this question by formalizing multi-rater labeling with a latent probability model in which each exam j has probability p_j of a positive label, and rater annotations are independent Bernoulli draws conditioned on p_j. Within this framework, we show that Expected Calibration Error (ECE) computed against majority vote violates the Oracle Guarantee: *even when a model predicts the true p_j, ECE does not vanish in expectation.* Our key contributions are:

- We formalize calibration under inter-rater variability and show that the non-linear majority-vote map induces transformations on the label distribution, fundamentally altering what it means to be calibrated.
- We investigate two principled alternatives: MR-ECE [24], which evaluates calibration directly against individual rater annotations, and AdjustedECE, which aligns model predictions with the consensus aggregation rule. We prove that both protocols recover the Oracle Guarantee asymptotically.
- Through controlled synthetic benchmarks and a real-world breast density case study, we empirically validate the theoretical insights.

Our work highlights that label aggregation can fundamentally change what it means to be calibrated and provides a principled framework for defining calibration targets that align with the intended interpretation of model predictions.

2 Background

2.1 Latent Probability Model for Multi-Rater Annotations

Medical imaging labels often exhibit inter-rater variability due to ambiguous visual evidence, imperfect image quality, heterogeneous diagnostic thresholds, and context-dependent clinical criteria [16,1,29]. Multi-rater studies render this stochasticity observable: each exam produces multiple annotations that can be viewed as samples from a case-specific labeling distribution. Standard evaluation collapses these into a single consensus label via majority vote.

Formally, consider an evaluation set of N_d exams indexed by $j = 1, \ldots, N_d$. Each exam has a latent single-rater positive probability $p_j \in [0, 1]$, and we observe n independent rater labels $y_{ij} \sim \text{Bernoulli}(p_j)$ for $i = 1, \ldots, n$. A case-level consensus label is constructed by majority vote, $\tilde{y}_j = \mathbb{1}\left[\sum_{i=1}^{n} y_{ij} \geq \left\lceil \frac{n}{2} \right\rceil\right]$. This formulation distinguishes two objects: (i) the *latent case uncertainty* p_j, and (ii) the *consensus label* $\tilde{y}_j$ produced by the aggregation protocol. The latter depends on the number of raters and varies with n even when p_j is fixed.

2.2 Evaluation Metrics for Measuring Calibration

Expected Calibration Error (ECE) [6] In standard settings, calibration is defined with respect to deterministic binary ground truth. Let a model output a predicted probability $\hat{p}_j$ for exam j with label $Y_j \in \{0, 1\}$. A predictor is

calibrated if $\Pr(Y_j = 1 \mid \hat{p}_j = p) = p$ for all $p \in [0, 1]$, that is, among instances assigned prediction p, the observed positive rate equals p. In practice, calibration is commonly measured by ECE $= \sum_{m=1}^{M} \frac{|B_m|}{N} |\text{frac}(B_m) - p_m|$, where predictions are partitioned into bins B_m, $\text{frac}(B_m)$ denotes the empirical fraction of ground truth positives in bin m, and p_m is corresponding mean prediction. ECE estimates the average deviation between predicted probabilities and observed frequencies. A fundamental consistency requirement, which we refer to as the *Oracle Guarantee*, is that if a predictor outputs the true conditional probability of the label, so that $\Pr(Y_j = 1 \mid \hat{p}_j = p) = p$ for all $p \in [0, 1]$, then calibration error vanishes in expectation: $\mathbb{E}[\text{frac}(B_m)] = p_m$ for all m, implying $\mathbb{E}[\text{ECE}] \to 0$.

Total Variation Distance (TVD) [3] Standard ECE assumes deterministic ground truth. In multi-rater settings, however, labels arise from the stochastic process $y_{ij} \sim \text{Bernoulli}(p_j)$. Baan et al. [3] show in multi-class settings that calibration measured against majority vote can be misleading under annotator disagreement. The same issue arises in our latent model: replacing rater labels with a consensus $\tilde{y}_j$ shifts the calibration target from p_j to a transformed aggregate, so even an oracle predictor with $\hat{p}_j = p_j$ may appear miscalibrated.

Under subjective labeling, calibration should instead be defined with respect to the distribution of rater judgments [3]. For the population target p_j, the average vote $\bar{y}_j = \frac{1}{n} \sum_{i=1}^{n} y_{ij}$ is an unbiased estimator. Thus, a natural instance-level measure is the total variation distance between the corresponding distributions, which reduces to $\text{TVD}_j = |\hat{p}_j - \bar{y}_j|$. Averaged across exams, TVD provides a principled reference for alignment between model predictions and human uncertainty. However, TVD requires multiple annotations per exam, whereas ECE is defined for a single label. In practice, clinical models are often developed using single-rater annotations but evaluated against multi-rater consensus, creating a mismatch between the development and testing environment.

2.3 Related Work

Extensive work has examined statistical limitations of ECE, including bias and instability from binning and finite samples [10,19,25,2,7], but under an implicit assumption of deterministic ground truth. In many real-world settings, labels are often stochastic due to inter-rater variability [4,22]. Related uncertainty-aware evaluation methods account for this through probabilistic labels [13], sampling-based evaluation [27], or metrics that compare model predictions against inter-rater disagreement [14], yet none directly address calibration.

In medical image segmentation, annotation uncertainty often arises from ambiguous boundaries [16]. Recent calibration evaluations therefore preserve rater disagreement rather than aggregating annotations [9,31]. The CURVAS challenge averages ECE computed separately for each expert annotation [23]. However, these methods do not explain how label aggregation changes the calibration target. Recent work has also proposed the multi-rater extension of calibration error (MR-ECE), which achieves greater stability compared to single-rater ECE [24].

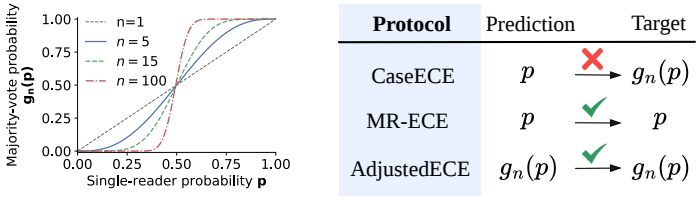

Fig. 1. Majority voting induces a nonlinear transformation $g_n(p)$ of the latent single-rater probability p, sharpening probabilities as n increases. (left). CaseECE compares p to $g_n(p)$, violating the Oracle Guarantee. MR-ECE targets p, while AdjustedECE aligns predictions with $g_n(p)$, restoring the Oracle Guarantee (right).

The method "virtually expands" the test set by treating per-rater annotations as individual samples within bins, framing its advantage in terms of variance reduction and robustness to limited test-set size. The emphasis is therefore on stability gains from multi-rater annotations. In contrast, we provide complementary theoretical characterization of MR-ECE under the lens of label aggregation. Specifically, we prove that MR-ECE aligns with the latent single-rater probability target, thus restoring the Oracle Guarantee under inter-rater variability.

Similar concerns arise in natural language processing, where annotation disagreement is pervasive [21] and its evaluation impact remains poorly understood [5]. Closest to our work, prior work shows that calibration against majority vote can be misleading and proposes TVD [3]. We instead show that ECE is valid under appropriate protocols. More broadly, handling annotation disagreement is an active and still-evolving area of research [30], reflected in dedicated workshops [12,28]. Recent evidence highlights that evaluation under aleatoric uncertainty should include calibration [11], as calibration measures alignment with the underlying stochastic process. Our work clarifies the calibration target and provides a theoretically grounded framework for valid evaluation in multi-rater settings.

3 How to Measure Calibration When Raters Disagree?

In multi-rater studies, calibration is typically evaluated against a single majority-vote label, which we denote as *CaseECE*. Majority voting induces a nonlinear map g_n of the single-rater probability, altering the calibration target and breaking the Oracle Guarantee (Fig. 1). We therefore consider two alternatives: *MR-ECE*, which evaluates calibration against individual rater annotations, and *AdjustedECE*, which aligns predictions with the consensus transformation.

Baseline: CaseECE. We analyze CaseECE under the latent multi-rater model introduced in § 2.1. Conditional on p_j, the vote count $S_j = \sum_{i=1}^{n} y_{ij}$ follows Binomial(n, p_j), so the majority-vote label $\tilde{y}_j$ satisfies

$$\mathbb{E}[\tilde{y}_j \mid p_j] = g_n(p_j), \qquad g_n(p) = \sum_{k=\lceil n/2 \rceil}^{n} \binom{n}{k} p^k (1-p)^{n-k}.$$

Thus, majority voting replaces the single-rater probability p_j with the transformed quantity $g_n(p_j)$. For $n = 1$, $g_1(p) = p$, but for $n > 1$ the mapping is nonlinear and increasingly steep around $p = 0.5$ as n grows (Fig. 1 (left)).

Recall that ECE compares, within each prediction bin B_m, the empirical positive fraction $\mathrm{frac}(B_m)$ to the bin-average prediction p_m (§ 2.2). For an oracle predictor $\hat{p}_j = p_j$, we have $p_m = \frac{1}{|B_m|}\sum_{j\in B_m} p_j$ and $\mathbb{E}[\mathrm{frac}(B_m) \mid \{p_j\}] = \frac{1}{|B_m|}\sum_{j\in B_m} g_n(p_j)$. Since g_n is nonlinear for $n > 1$, these quantities generally differ. As the number of exams per bin increases and empirical fractions concentrate around their expectations, CaseECE converges to a nonzero limit. Hence, even an oracle predictor does not achieve zero calibration error in expectation. For notational simplicity, we assume a fixed rater count n per exam; the formulation extends directly to case-specific counts n_j that vary by exam.

Remedy I: MR-ECE. The Oracle Guarantee is restored if calibration is computed against individual rater labels, without applying majority-vote aggregation. *MR-ECE* is defined as ECE computed by using all individual rater labels within each prediction bin [24]. The empirical positive fraction becomes:

$$\mathrm{frac}_{\mathrm{rater}}(B_m) = \frac{1}{|B_m|\, n} \sum_{j\in B_m} \sum_{i=1}^{n} y_{ij}.$$

For an oracle predictor $\hat{p}_j = p_j$,

$$\mathbb{E}[\mathrm{frac}_{\mathrm{rater}}(B_m) \mid \{p_j\}] = \frac{1}{|B_m|} \sum_{j\in B_m} p_j = p_m.$$

As the total number of rater annotations per bin increases, the empirical fraction concentrates around its expectation, and MR-ECE converges to zero asymptotically. MR-ECE thus evaluates calibration with respect to the single-rater probability p_j, preserving the Oracle Guarantee in expectation.

Remedy II: AdjustedECE. An alternative solution is to retain consensus (majority-vote) labels and measure calibration explicitly with respect to this transformed target. Instead of removing the aggregation, we apply the same majority-vote map to the predictions. Starting from an oracle single-rater predictor $\hat{p}_j = p_j$, define $\hat{q}_j = g_n(\hat{p}_j) = g_n(p_j)$. ECE is then computed between $\hat{q}_j$ and the consensus labels $\tilde{y}_j$. For a bin B_m defined using $\hat{q}_j$,

$$\mathbb{E}\left[\frac{1}{|B_m|}\sum_{j\in B_m} \tilde{y}_j \;\middle|\; \{p_j\}\right] = \frac{1}{|B_m|}\sum_{j\in B_m} g_n(p_j) = \frac{1}{|B_m|}\sum_{j\in B_m} \hat{q}_j.$$

Thus, after passing predictions through the same transformation g_n, the bin-average prediction equals the expected consensus rate. Since g_n is continuous, AdjustedECE converges to zero for the transformed oracle predictor. AdjustedECE therefore evaluates calibration with respect to the majority-vote probability $g_n(p_j)$, preserving the Oracle Guarantee under consensus labeling.

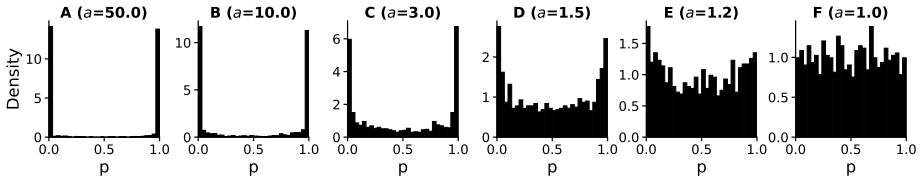

Fig. 2. Distributions of latent probabilities p across synthetic worlds A–F. Each world is defined by a concentration parameter a (indicated in the panel title), controlling the degree of subjectivity; worlds A–F are ordered from lower to higher subjectivity.

4 Benchmark Experiment

To test the Oracle Guarantee under inter-rater variability, we construct a synthetic benchmark with known latent probabilities p_j. Using oracle predictions $\hat{p}_j = p_j$, we examine the asymptotic behavior of different calibration protocols. [1]

4.1 Constructing the synthetic benchmark

To simulate varying degrees of subjectivity, we generated symmetric bimodal probability distributions via monotone power transformations of Uniform$(0, 1)$ random variables. Let $U \sim \text{Uniform}(0, 1)$. We set $p = 0.5\,U^a$ for half of the cases and $p = 1 - 0.5\,U^a$ for the remainder. The parameter $a > 0$ controls concentration: larger a places more mass near 0 and 1 (higher objectivity), while smaller a shifts mass toward 0.5 (greater subjectivity). Fig. 2 shows the resulting distributions (Worlds A–F), illustrating a monotonic progression toward increasing intrinsic subjectivity. As a decreases, a larger fraction of cases concentrate near $p = 0.5$, increasing expected inter-rater disagreement.

On realism and applicability. The symmetric generative scheme is not intended to uniquely model rater subjectivity; alternative constructions could similarly span regimes from near-deterministic labeling to highly subjective cases. Our objective is not to replicate all aspects of real-world rater behavior, but to provide a controlled and interpretable mechanism, governed by a single parameter a, for varying intrinsic ambiguity and testing whether MR-ECE and AdjustedECE can recover the Oracle Guarantee across varying subjectivity levels.

Simulation protocol. For each world (A–F), we draw latent probabilities $\{p_j\}_{j=1}^{N_d}$ for N_d $(= 1000)$ exams from the corresponding world distribution, and generate rater labels $y_{ij} \mid p_j \sim \text{Bernoulli}(p_j)$. We evaluate an oracle predictor that outputs the true latent probabilities, $\hat{p}_j = p_j$. Metrics are computed by comparing $\hat{p}_j$ to labels derived from rater annotations using either (i) individual rater-level labels $\{y_{ij}\}$ or (ii) majority vote $\tilde{y}_j$, depending on the evaluation protocol. This experiment tests whether each protocol satisfies the Oracle Guarantee, namely convergence to zero under oracle predictions.

[1] `https://github.com/whiterabbit-ai/calibration-under-rater-disagreement`.

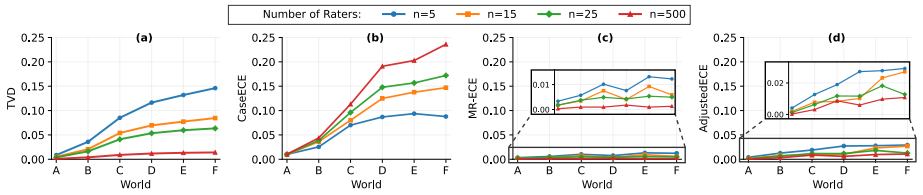

Fig. 3. Effect of rater sample size and task subjectivity on calibration-related quantities across synthetic worlds (A–F). (a) TVD, (b) CaseECE, (c) MR-ECE, and (d) AdjustedECE. MR-ECE and AdjustedECE track TVD, whereas CaseECE deviates.

4.2 Results

We report results for $n = 5, 15, 25, 500$ raters per exam in Fig. 3. Larger values of n allow us to probe the asymptotic behavior of the evaluation metrics. As n increases, sampling variability decreases, so observed trends more closely reflect the intrinsic uncertainty of the task, which remains unchanged.

Panel (a) shows TVD. For small n, TVD is inflated by sampling variability in estimating $\bar{y}_j$. As n increases, TVD decreases toward 0, exhibiting the expected asymptotic convergence. This confirms TVD as a suitable reference metric.

Panel (b) shows CaseECE. Unlike TVD, CaseECE increases with larger rater counts. This opposite trend reflects the shift in calibration target induced by majority voting and confirms that CaseECE violates the Oracle Guarantee.

Panels (c) and (d) show MR-ECE and AdjustedECE. Both approach zero at high rater counts and remain relatively stable across subjectivity levels, confirming preservation of the Oracle Guarantee. Their trends mirror TVD as n increases. MR-ECE is more sample-efficient: since each case contributes n annotations, bins contain more samples, leading to faster convergence.

5 Case Study: Calibration Performance of Density Model

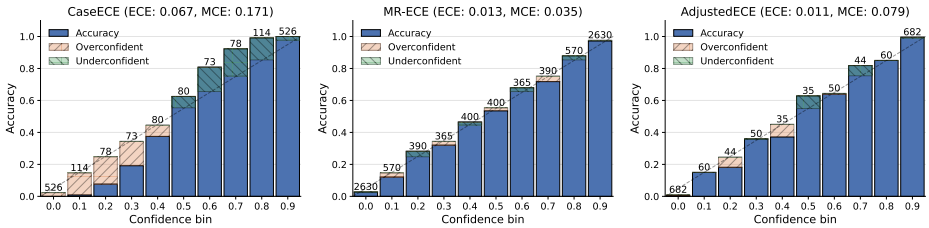

Fig. 4. Calibration curves of a breast density model for CaseECE (left), MR-ECE (middle), and AdjustedECE (right). Each panel reports ECE and Maximum Calibration Error (MCE). Bars show empirical accuracy per bin, with counts indicated above.

We evaluate a breast density model [32,15] on the binary classification task of high versus low breast density. Due to the cost of multi-rater labels, model development and calibration relied exclusively on single-rater labels. At test time, however, the model is evaluated on a dataset with five raters per exam. Fig. 4 shows the resulting calibration curves under the three evaluation protocols. The model appears miscalibrated with CaseECE of 0.067 (95%CI: 0.054, 0.081), and the bin patterns closely resemble $g_5(p)$ in Fig. 1, as expected by our theoretical framework. When computed against individual raters, the model has ECEs of 0.029 (0.023, 0.049), 0.033 (0.018, 0.055), 0.044 (0.032, 0.060), 0.053 (0.040, 0.068), 0.095 (0.075, 0.123), i.e., less than CaseECE for 4/5 raters and significantly less for the first three raters ($p < 0.005$ using 10,000 paired bootstraps). These findings support that CaseECE suffers from target mismatch. Meanwhile, MR-ECE and AdjustedECE align closely with the ideal calibration line. This discrepancy has important regulatory implications, illustrating how target mismatches affect perceived performance and conclusions about model acceptability.

6 Discussion

Practical Considerations. Tab. 1 summarizes the trade-offs between MR-ECE and AdjustedECE. *(1) Sample efficiency:* With N_d exams and n raters per exam, TVD and AdjustedECE operate on N_d exam-level labels. MR-ECE utilizes all $N_d n$ rater-level labels, thus improving statistical stability through larger effective sample size [24]. *(2) Label requirements and aggregation assumption:* MR-ECE is preferable when rater-level labels are available because it does not require an aggregation model. AdjustedECE relies on the aggregation map g_n and is most useful when only consensus labels are available, as is common in medical datasets [8,18]. *(3) Reliability-diagram resolution:* MR-ECE retains resolution across the full probability range $p \in [0, 1]$. In contrast, g_n pushes scores toward 0 and 1 as the number of raters grows, reducing the informativeness of AdjustedECE reliability diagrams (Fig. 4). For sufficiently large n, this can yield low AdjustedECE even for miscalibrated models because most samples occupy extreme bins. In practice, the effect is limited by the small number of raters in most medical datasets and can be further mitigated by adaptive binning [19].

The i.i.d. assumption. We adopt i.i.d. raters as a mathematically clean setting for isolating calibration-target mismatch. Empirically, the breast density study supports the simplified i.i.d. model, suggesting that it captures the dominant

Table 1. MR-ECE vs. AdjustedECE.

Metric	Sample Count	Labels Required	Aggregation Model	Reliability Diagram Resolution
MR-ECE	$N_d n$	Rater-level	None	Stable with increasing raters
AdjustedECE	N_d	Consensus only	Required	Reduces with increasing raters

effects of multi-rater aggregation in practice. Further, relaxing the i.i.d. assumption changes the aggregation map but not the calibration-target mismatch. For example, if rater-specific probabilities p_j^i vary across raters i, majority voting induces a Poisson-binomial aggregation map h_n rather than g_n. MR-ECE then targets the mean rater probability, $\frac{1}{n}\sum_{i=1}^{n} p_j^i$, while AdjustedECE applies h_n to align predictions with the consensus target. Thus, relaxing i.i.d. assumption replaces the closed-form map g_n with a more general operator.

Conclusion. The apparent failure of calibration in subjective settings stems from misalignment between model predictions and the label aggregation rule. MR-ECE and AdjustedECE restore consistency by aligning calibration with the correct probabilistic target. Our results establish that calibration remains valid under label stochasticity when properly defined, laying a principled foundation for evaluating calibration error in subjective tasks.

While we focus on binary classification, the framework naturally extends to several important settings, which could be explored through future work. These include multiclass classification, where aggregation may similarly alter calibration targets; annotation processes with heterogeneous or correlated raters, which require new aggregation maps; and other prediction tasks that involve stochastic labels, such as medical image segmentation.

Disclosure of Interests. All authors are Whiterabbit.ai employees and may hold equity. Whiterabbit.ai developed the breast density model evaluated in this study.

References

1. Alomaim, W., O'Leary, D., Ryan, J., Rainford, L., Evanoff, M., Foley, S.: Variability of breast density classification between us and uk radiologists. Journal of Medical Imaging and Radiation Sciences **50**(1), 53–61 (2019)
2. Arrieta-Ibarra, I., Gujral, P., Tannen, J., Tygert, M., Xu, C.: Metrics of calibration for probabilistic predictions. JMLR **23**(351), 1–54 (2022)
3. Baan, J., Aziz, W., Plank, B., Fernandez, R.: Stop measuring calibration when humans disagree. In: Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing. pp. 1892–1915. Association for Computational Linguistics, Abu Dhabi, United Arab Emirates (Dec 2022)
4. Chen, P.H.C., et al.: Evaluation of artificial intelligence on a reference standard based on subjective interpretation. The Lancet Digital Health **3**(11) (2021)
5. Gordon, M.L., et al.: The disagreement deconvolution: Bringing machine learning performance metrics in line with reality. In: Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems (2021)
6. Guo, C., Pleiss, G., Sun, Y., Weinberger, K.Q.: On calibration of modern neural networks. In: Proceedings of the 34th International Conference on Machine Learning. pp. 1321–1330. PMLR (2017)
7. Gupta, K., Rahimi, A., Ajanthan, T., Sminchisescu, C., Mensink, T., Hartley, R.I.: Calibration of neural networks using splines. In: International Conference on Learning Representations (ICLR) (2021)

8. Irvin, J., et al.: Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison. In: Proceedings of the AAAI conference on artificial intelligence. pp. 590–597 (2019)
9. Ji, W., et al.: Learning calibrated medical image segmentation via multi-rater agreement modeling. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 12336–12346 (2021)
10. Kumar, A., Sarawagi, S., Jain, U.: Trainable calibration measures for neural networks from kernel mean embeddings. In: Proceedings of the 35th International Conference on Machine Learning. vol. 80, pp. 2805–2814. PMLR (2018)
11. Kumar, H., Kang, B., Chakraborty, B., Mukhopadhyay, S.: Has the deep neural network learned the stochastic process? an evaluation viewpoint. In: Proceedings of the International Conference on Learning Representations (ICLR) (2025)
12. Leonardelli, E., et al.: SemEval-2023 task 11: Learning with disagreements (LeWiDi). In: Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023). pp. 2304–2318. Association for Computational Linguistics, Toronto, Canada (Jul 2023)
13. Lionetti, S., Gröger, F., Gottfrois, P., Gonzalez-Jimenez, A., Amruthalingam, L., Navarini, A.A., Pouly, M.: Clinical uncertainty impacts machine learning evaluations. In: Machine Learning for Health 2025 (2025)
14. Lovchinsky, I., Daks, A., Malkin, I., Samangouei, P., Saeedi, A., Liu, Y., Sankaranarayanan, S., Gafner, T., Sternlieb, B., Maher, P., Silberman, N.: Discrepancy ratio: Evaluating model performance when even experts disagree on the truth. In: International Conference on Learning Representations (2020)
15. Matthews, T.P., et al.: A multisite study of a breast density deep learning model for full-field digital mammography and synthetic mammography. Radiology: Artificial Intelligence **3**(1), e200015 (2020)
16. Monteiro, M., Le Folgoc, L., Coelho de Castro, D., Pawlowski, N., Marques, B., Kamnitsas, K., Van der Wilk, M., Glocker, B.: Stochastic segmentation networks: Modelling spatially correlated aleatoric uncertainty. Advances in neural information processing systems **33**, 12756–12767 (2020)
17. Naeini, M.P., Cooper, G., Hauskrecht, M.: Obtaining well calibrated probabilities using bayesian binning. In: Proceedings of the AAAI conference on artificial intelligence. vol. 29 (2015)
18. Nguyen, H.Q., Lam, K., Le, L.T., Pham, H.H., Tran, D.Q., Nguyen, D.B., Le, D.D., Pham, C.M., Tong, H.T., Dinh, D.H., et al.: Vindr-cxr: An open dataset of chest x-rays with radiologist's annotations. Scientific Data **9**(1), 429 (2022)
19. Nixon, J., Dusenberry, M.W., Zhang, L., Jerfel, G., Tran, D.: Measuring calibration in deep learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops (2019)
20. Panayides, A.S., et al.: Ai in medical imaging informatics: Current challenges and future directions. IEEE Journal of Biomedical and Health Informatics **24**(7), 1837–1857 (2020). https://doi.org/10.1109/JBHI.2020.2991043
21. Pavlick, E., Kwiatkowski, T.: Inherent disagreements in human textual inferences. Transactions of the Association for Computational Linguistics **7**, 677–694 (2019). https://doi.org/10.1162/tacl_a_00293
22. Reinke, A., et al.: Understanding metric-related pitfalls in image analysis validation. Nature methods **21**(2), 182–194 (2024)
23. Riera-Marín, M., Sikha, O., Rodríguez-Comas, J., May, M.S., Pan, Z., Zhou, X., Liang, X., Erick, F.X., Prenner, A., Hémon, C., et al.: Calibration and uncertainty for multirater volume assessment in multiorgan segmentation (curvas) challenge results. Computers in Biology and Medicine **197**, 111024 (2025)

24. Riera-Marín, M., et al.: Multi-rater calibration error estimation. In: International Workshop on Uncertainty for Safe Utilization of Machine Learning in Medical Imaging. pp. 147–157. Springer (2025)
25. Roelofs, R., et al.: Mitigating bias in calibration error estimation. In: Proceedings of The 25th International Conference on Artificial Intelligence and Statistics. Proceedings of Machine Learning Research, vol. 151, pp. 4036–4054 (2022)
26. Salari, A., Djavadifar, A., Liu, X., Najjaran, H.: Object recognition datasets and challenges: A review. Neurocomputing **495**, 129–152 (2022)
27. Stutz, D., et al.: Evaluating medical ai systems in dermatology under uncertain ground truth. Medical Image Analysis **103**, 103556 (2025)
28. Sudre, C.H., Hoque, M.I., Mehta, R., Ouyang, C., Qin, C., Rakic, M., Wells, W.M. (eds.): Uncertainty for Safe Utilization of Machine Learning in Medical Imaging, Lecture Notes in Computer Science, vol. 16166. Springer Cham (2025). https://doi.org/10.1007/978-3-032-06593-3
29. Théberge, I., Guertin, M.H., Vandal, N., Daigle, J.M., Dufresne, M.P., Wadden, N., Shumak, R., Samson, C., Langlois, A., Larocque, I., et al.: Clinical image quality and sensitivity in an organized mammography screening program. Canadian Association of Radiologists Journal **69**(1), 16–23 (2018)
30. Uma, A.N., Fornaciari, T., Hovy, D., Paun, S., Plank, B., Poesio, M.: Learning from disagreement: A survey. Journal of Artificial Intelligence Research **72** (2021)
31. Wang, X., Yang, M., Tosun, S., Nakamura, K., Li, S., Li, X.: Caldiff: Calibrating uncertainty and accessing reliability of diffusion models for trustworthy lesion segmentation. IEEE Journal of Biomedical and Health Informatics **30**(2), 1555–1567 (2026). https://doi.org/10.1109/JBHI.2025.3624331
32. Whiterabbit.ai Inc.: WRDensity by Whiterabbit.ai: 510(k) Premarket Notification Summary. 510(k) Premarket Notification Summary K202013, U.S. Food and Drug Administration (2020), https://www.accessdata.fda.gov/scripts/cdrh/cfdocs/cfpmn/pmn.cfm?ID=K202013