On Evaluating Subgroup Discovery in Medical Image Classification

Alceu Bissoto[1,2,3], Trung-Dung Hoang[1,2,3], Tim Flühmann[1,2,3], Emma A. M. Stanley[4], Raghav Mehta[4], Ben Glocker[4], and Lisa M. Koch[1,2,3]

[1] University of Bern, Switzerland
[2] Department of Diabetes, Endocrinology, Nutritional Medicine and Metabolism, Bern University Hospital, Switzerland
[3] Diabetes Center Berne, Switzerland
[4] Imperial College London, United Kingdom
{alceu.bissoto,lisa.koch}@unibe.ch

Abstract. Subgroup annotations are central to medical AI applications (e.g., for performance monitoring and robust training) but remain scarce in most medical datasets. Subgroup discovery methods (SDMs) offer a promising alternative by learning subgroup labels from data representations. However, current evaluation strategies limit their applicability by relying on target ground-truth labels, mainly in toy setups with artificially induced groups. We propose an evaluation framework for SDMs in medical imaging classification to benchmark state-of-the-art SDMs across five medical imaging tasks: performance gap identification, deferral, robustness to random seeds, metadata alignment, and domain adaptation. Our evaluation framework reveals that simple K-Means clustering matches or outperforms sophisticated methods in most settings, and that SDMs are highly sensitive to hyperparameters and random initialization, requiring adaptation for real-world deployment. This work provides the first practical benchmark for subgroup discovery in medical imaging, offering evidence-based guidance for method selection while highlighting critical gaps for future research.[5]

Keywords: subgroup discovery · trustworthy AI · post-market surveillance.

1 Introduction

Subgroup annotations can enable a wide range of tasks in medical AI deployment. Subgroup labels allow practitioners to audit models for performance disparities [2], guide targeted improvements through domain adaptation [1], monitor distribution shifts across time [18], and implement selective prediction through deferral mechanisms [23]. Metadata attributes, such as patient age or sex, are typically used as subgroup annotations, but rich metadata are rarely available and may not adequately capture meaningful variability in the data. On the other

[5] Code: https://github.com/alceubissoto/divisions_benchmark

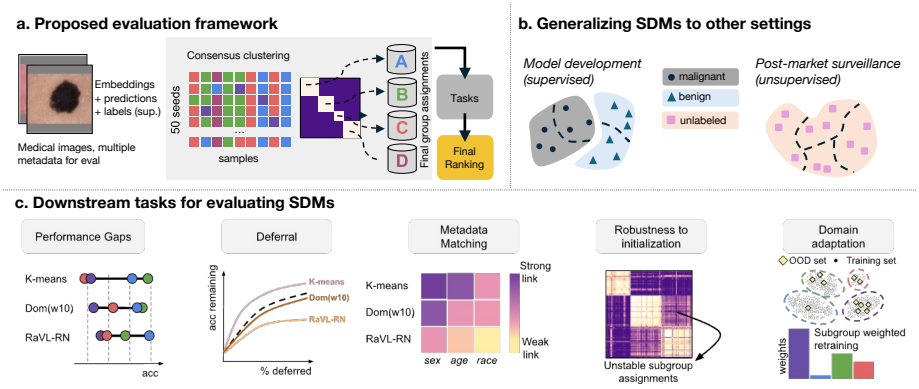

Fig. 1. a) We propose a robust evaluation framework for subgroup discovery across five downstream tasks. b) We evaluate SDMs in both supervised and unsupervised settings. c) Our proposed tasks capture diverse subgroup characteristics.

hand, subgroup discovery methods (SDMs) [7,24,15,21,22] aim to reveal hidden stratifications [14] by partitioning datasets into coherent groups based on model representations and error patterns. However, existing methods have been evaluated primarily for a narrow application: identifying known failure modes in controlled experimental settings.

This narrow focus creates significant limitations for practical deployment: i) SDMs' reliance on ground truth labels for bias analysis, and ii) their incomplete evaluation. First, most SDMs require access to task labels (e.g., disease diagnoses) at test time to assign samples to discovered subgroups [7,24,15]. This requirement restricts their use in critical post-deployment scenarios such as distribution shift monitoring or out-of-distribution detection, where test labels are unavailable (Fig. 1b). Second, evaluations have centered on retrieving predefined biases in synthetic settings with known ground-truth subgroups [7,21]. The standard metric, Precision-at-k [7], assesses only the top-k (usually 10) samples from the best-performing subgroup, ignoring partition quality across the full dataset.

We propose a practical evaluation framework for subgroup discovery methods and use it to benchmark existing methods that utilize a variety of feature extraction and clustering techniques (Fig. 1). We introduce minimal but necessary adaptations to make existing SDMs deployable without ground truth task labels, and benchmark them alongside their original formulation and a simple baseline (`K-Means`) across two medical imaging datasets (CheXpertPlus [4] and HAM10000 [20]) and five downstream tasks spanning performance monitoring, selective prediction, robustness to random initialization, interpretability, and domain adaptation.

Along with our framework, we contribute to the field by (1) demonstrating that a simple baseline can match state-of-the-art SDMs, and (2) identifying and mitigating the overlooked issue of SDM instability across random initializations. These contributions bridge the gap between subgroup discovery and medical

imaging research, enabling developers to benchmark new methods and helping practitioners select the best approach for their application context.

2 Methods

We propose an evaluation framework for SDMs and benchmark existing state-of-the-art methods. For each dataset, we train a disease classification model, fit SDMs on the validation set, and predict subgroup assignments across train and test. Because all SDMs have stochastic components, we repeat each assignment across 50 random seeds and aggregate via consensus clustering [13] (Sec. 2.2), substantially improving assignment stability. We evaluate the resulting partitions on five downstream tasks that comprise our comprehensive evaluation framework (Sec. 2.3). Finally, we aggregate rankings across tasks and datasets using the Bradley-Terry [3], which has gained popularity for ranking language models on the LMArena platform [5]. The Bradley-Terry model estimates a latent strength parameter per method from pairwise comparisons, yielding a global ranking that is more robust to inconsistencies than simple average rank aggregation.

2.1 Subgroup Discovery Methods

Most SDMs rely on three information sources: feature embeddings, model predictions, and ground truth labels. To extend SDMs to post-deployment monitoring tasks, we introduce two deployment-oriented properties: (1) no reliance on test-time labels for the classification task, and (2) complete assignment of every sample to at least one subgroup. We introduce minimal changes to adapt each evaluated SDM to meet the above criteria ("unsupervised" setting). Then, we benchmark them against their original design ("supervised", using ground-truth task labels). Whenever applicable, we also contrast SDMs with confidence- and metadata-based baselines.

For **K-Means**, we extract penultimate-layer embeddings from the target classification model, apply PCA to reduce their dimensionality to 128, and apply K-Means clustering. **DOMINO (DOM) [7]** fits a Gaussian Mixture Model (GMM) over a combined representation of external embeddings (e.g., CLIP [16], BioMedCLIP [25]) and target model predictions, weighted by a hyperparameter w that increases the relative influence of predictions over embeddings. Originally, `DOM` used ground-truth task labels to train the GMM, but we removed this reliance as proposed in [2]. **RaVL [21]** operates at the patch level. Images are first divided into a grid, and embeddings (using CLIP or the target model) are computed for each patch. Then, they fit K-Medoids clustering to patch representations, yielding patch-level subgroup assignments per sample. To assign each image to a single subgroup, we propose aggregating patch-level assignments using majority voting. **CoSi [24]** follows a similar structure to `DOM`, but trains a bias-amplified classifier to provide predictions. **STB [15]** also follows `DOM`'s structure, but adds a fully connected layer after the penultimate layer of the target model for dimensionality reduction. Both `CoSi` and `STB` fit a separate

GMM per class in their supervised versions, resulting in K groups per class. We modify them to fit a single GMM instead in their unsupervised versions.

We exclude LLM-guided approaches [9,8] from our analysis as they rely on domain-specific textual resources that are not consistently available in medical imaging and cannot capture hidden stratifications absent from text corpora.

2.2 Consensus Clustering

All SDMs we benchmark are sensitive to random initialization. To address this, we adopt consensus clustering [13]. Each SDM is run 50 times with different random seeds; each run fits the model on 80% of the validation set and generates predictions for all splits. To derive robust consensus subgroup assignments, we compute a normalized consensus matrix $M_{i,j}$ whose entries record the frequency of pairwise co-occurrence across runs, and apply agglomerative hierarchical clustering. We cut the resulting dendrogram at K clusters. Finally, subgroup labels are propagated to the test set using a 1-Nearest Neighbor classifier with the Hamming distance, mapping each new sample's co-occurrence profile to the most similar membership history in the validation set. This procedure stabilizes subgroup assignments and provides a natural basis for stability evaluation, one of our downstream tasks to evaluate SDMs (Sec. 2.3).

2.3 Evaluation Tasks

We propose evaluating SDMs using five downstream tasks that reflect distinct practical needs.

Performance gap identification. A common use case of subgroup annotations is to reveal systematic performance disparities of a classification model [19,12]. For each subgroup partition, we compute the classification accuracy within each subgroup and report the difference between the top and bottom 20th percentile groups. A larger gap indicates that the partition successfully separates high- and low-performing regions of the data. As baselines, we contrast SDMs with purely confidence-based and metadata-based divisions.

Selective prediction (deferral). In safety-critical applications, it may be preferable to defer predictions on samples where the model is error-prone [11]. For all SDMs, we defer samples from the lowest-validation-performance subgroups first, breaking ties by individual prediction confidence, and report accuracy on the retained 80% after deferring the bottom 20%. Similar to performance gaps, we include confidence-based divisions as a baseline.

Robustness to random initialization. A subgroup partition is only useful in practice if it is reproducible. We measure robustness via the consensus matrix from 50 runs (see Sec. 2.2), which records how often each pair of samples is assigned to the same subgroup. We report the gap between within-group and between-group average co-occurrence, where higher values indicate that the method consistently groups the same samples across initializations.

Metadata alignment. Subgroups are more actionable if they align with interpretable characteristics. We assess whether discovered subgroups align with

clinical metadata by fitting logistic regression models of the form (metadata $\sim$ subgroup + target). The coefficient associated with subgroup membership is the log-odds ratio between subgroup assignment and metadata after adjusting for the target label. The log-odds inform how much more likely the metadata is to appear in the subgroup. We report a coverage metric, where we weight each metadata attribute by its associated risk and consider it captured by an SDM if the odds ratio exceeds 2.0 and survives false discovery rate correction at $\alpha = 0.05$.

Domain adaptation. Subgroup annotations can support domain adaptation. We first assign subgroup labels to out-of-distribution (OOD) samples using the fitted SDM, then compare the subgroup prevalence between in-distribution and OOD data, and finally retrain the classifier with importance weights reflecting the OOD subgroup distribution. We report the OOD AUC after reweighted training, averaged over three runs. The aim of this downstream task is to rank SDMs by their ability to capture distribution-relevant structure, rather than to outperform state-of-the-art domain adaptation.

2.4 Datasets and classification tasks

We evaluate SDMs on two medical image classification tasks:

CheXpertPlus [4] is a chest X-ray dataset of 223,228 images (178,684 training, 22,263 validation, 22,281 test) with multi-label annotations for 14 radiological findings. We focus on cardiomegaly detection as a classification task. The dataset provides metadata, including patient demographics, comorbidities, and support devices. For domain adaptation, we use a subset of BRAX [17].

HAM10000 [20] is a dermoscopic image dataset for skin lesion classification across seven labels with 10,015 samples. Our task of choice is melanoma *vs.* benign, thus we remove carcinomas from the dataset, resulting in 5,762 training, 1,499 validation, and 1,822 test images. The dataset provides limited metadata (lesion location, patient age, sex, and diagnostic method). To enable richer metadata alignment evaluation, we construct and make available additional annotations by applying ground truth lesion segmentation masks and extracting lesion colour and shape features inspired by the metadata of SLICE-3D [10]. For domain adaptation, we use a subset of BCN20000 [6].

3 Results

3.1 Experimental setup

We set the number of subgroups to $K = 15$ for CheXpertPlus, and $K = 10$ for HAM10000 to account for its smaller size. Fixing K across all SDMs, rather than allowing methods to tune it, ensures a fair comparison. We used a ResNet-50 as our classification model. For `DOM`, we evaluated three weight settings ($w \in \{1, 10, 100\}$) and two external embedding models (CLIP [16] and Biomed-CLIP [25]), treating each combination as a separate variant (`DOM` and `DOM-Bio`, respectively). For `RaVL`, we evaluated patches extracted in a 4×4 grid and used CLIP (`RaVL-CLIP`) and the target model (`RaVL-RN`) to provide embeddings. For `CoSi` and `STB`, we used the original hyperparameter search and values.

Table 1. Comparison of subgroup division methods across five tasks: Performance gap identification (Gap), Deferral (Def), Metadata alignment (Met), Stability to random initialization (Rob), domain adaptation (OOD). We use a Bradley–Terry (BT) model to aggregate scores into a final ranking. We highlight best and second-best results. Deferral is not available in the supervised setting (–).

	HAM10000					CheXpert Plus					
Method	**Gap**	**Def**	**Met**	**Rob**	**OOD**	**Gap**	**Def**	**Met**	**Rob**	**OOD**	**BT**
Unsupervised											
`K-Means`	58.8	76.5	91.7	65.8	70.1	66.2	88.3	72.6	92.8	81.9	2.71
`DOM(w100)`	63.8	80.2	86.2	88.9	65.4	61.5	89.2	16.2	74.8	81.8	1.16
`DOM(w10)`	67.1	76.0	86.7	50.3	69.4	55.3	87.7	56.0	77.8	83.2	1.65
`DOM(w1)`	44.4	72.4	82.8	42.6	65.6	10.8	79.8	61.8	64.2	81.0	0.14
`DOM-Bio(w100)`	64.9	80.1	83.9	78.6	67.8	60.8	87.6	50.9	81.2	82.3	1.51
`DOM-Bio(w10)`	61.2	78.3	84.5	51.6	70.2	51.5	87.0	60.4	70.0	81.2	0.90
`DOM-Bio(w1)`	48.6	74.8	80.1	49.0	68.4	23.2	80.8	61.9	70.8	82.1	0.40
`STB`	50.8	73.7	86.9	72.1	66.6	58.6	86.5	62.0	64.3	84.5	0.94
`CoSi`	65.6	77.1	81.3	97.2	68.8	51.2	84.8	37.5	80.9	81.9	0.90
`RaVL-CLIP`	47.4	74.5	90.2	50.3	68.7	7.3	78.7	40.9	12.8	82.9	0.30
`RaVL-RN`	54.1	74.8	86.2	75.6	65.3	13.0	79.4	43.6	12.9	83.2	0.40
Supervised											
`K-Means` (unsup.)	58.8	–	91.7	65.8	70.1	66.2	–	72.6	92.8	81.9	1.98
`DOM(w100)`	99.1	–	89.1	91.8	67.4	99.5	–	27.2	84.4	81.4	1.67
`DOM(w10)`	93.7	–	88.3	57.3	67.9	72.8	–	54.6	70.2	81.8	1.09
`DOM(w1)`	49.6	–	86.4	41.1	69.3	11.2	–	55.8	64.7	82.4	0.65
`DOM-Bio(w100)`	99.2	–	84.8	86.5	63.1	99.3	–	42.3	81.6	81.4	0.84
`DOM-Bio(w10)`	89.0	–	85.0	65.7	67.3	71.5	–	62.3	74.9	81.4	0.80
`DOM-Bio(w1)`	47.7	–	85.4	50.3	66.0	22.8	–	60.6	68.4	82.2	0.56
`STB`	75.1	–	83.0	70.4	56.4	78.6	–	63.9	63.5	82.1	0.69
`CoSi`	91.3	–	91.6	94.6	67.1	71.9	–	43.1	88.9	83.0	2.23
`RaVL-CLIP`	41.0	–	78.9	47.5	65.6	5.9	–	39.9	52.9	82.7	0.22
`RaVL-RN`	42.0	–	78.4	73.2	65.5	0.3	–	22.0	30.8	84.0	0.27

3.2 SDMs performance varies across downstream tasks

Target model guidance helps to identify high-performance gaps. The models that revealed the highest accuracy gaps (above 55% in both datasets for the unsupervised setting) were strongly influenced by the target model, e.g. `DOM` variants that put higher weight (`w >= 10`) on predictions, or `K-means`, which cluster target model features directly (Tab. 1, "Gap" columns). Unsurprisingly, access to ground-truth labels improves results, as SDMs directly exploit model errors. Yet, the gaps found in the unsupervised setting far exceeded disparities in standard metadata (44.0% and 11.4% for HAM1000 and CheXpertPlus, respectively) and in pure confidence-based subgroups (46.1% and 44.0%), confirming that SDMs offer a meaningful advantage for performance monitoring [2].

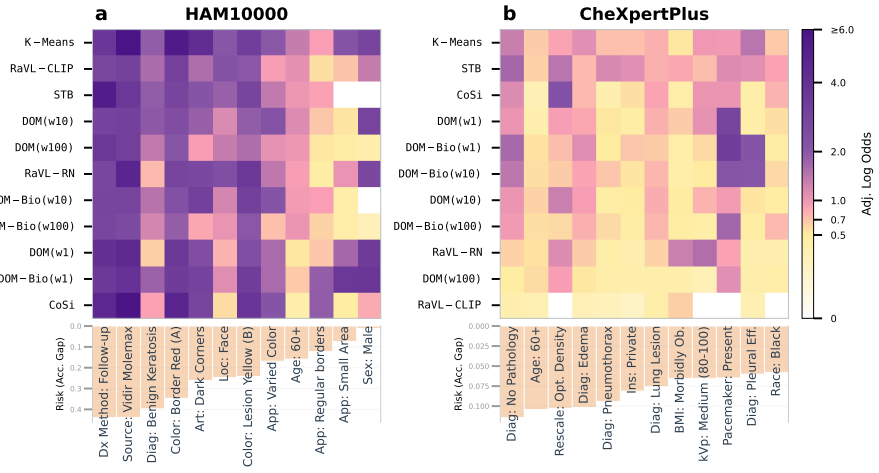

Fig. 2. Metadata matching results for the unsupervised setting. Heatmap cells represent log-odds ratios. Rows are sorted by the best methods for each dataset, and columns are sorted by the metadata that cause the highest accuracy gaps.

SDMs with target model guidance outperform confidence-based deferral. Similar SDMs also performed best in the deferral task (Tab. 1, "Def" columns), with `DOM` variants (`w` $\geq$ 10), `K-Means` outperforming the baseline (where samples were deferred based on confidence alone) and other SDMs on both datasets. The deferral setting clarifies why confidence alone (73.3% and 82.7% for HAM1000 and CheXpertPlus, respectively) is insufficient: it cannot distinguish high-confidence errors from low-confidence correct predictions. Incorporating embeddings (as in `DOM`) encourages clustering samples according to semantic concepts, providing context to defer problematic high-confidence cases.

SDMs capture high-risk attributes. `K-Means` identified subgroups most aligned with high-risk attributes, capturing 91.7% and 72.6% of the risk associated with metadata for HAM10000 and CheXpertPlus, respectively (Tab. 1, "Met" columns), surpassing all SDMs even in the supervised setting. Most other SDMs also demonstrated strong alignment with high-disparity metadata (Fig. 2), with log-odds ratios typically above 1 across both datasets (*i.e.*, attribute was at least 2.7 times more likely to appear in the identified subgroup). The inclusion of ground-truth labels (supervised setting) does not seem to affect this task much. Results were consistently stronger on HAM10000, where metadata focused on more easily encoded color and shape features, than on CheXpertPlus, where complex comorbidities and patient characteristics posed a harder challenge.

SDMs are sensitive to random seeds. All compared SDMs were sensitive to random seeds, so samples were often not consistently assigned together across runs in both settings. Robustness varied significantly between SDMs (Fig. 3 and Tab. 1, "Rob" columns). While some configurations led to robust partitions (e.g.

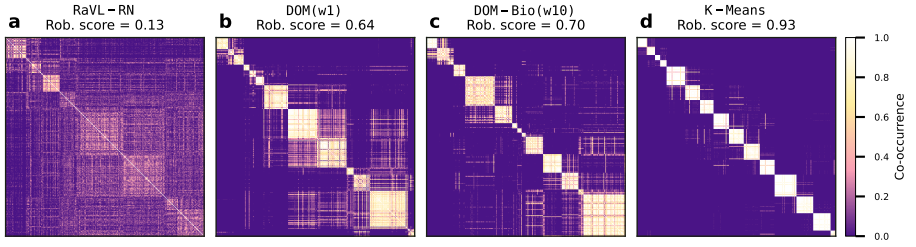

Fig. 3. Consensus matrices showing the co-occurrence of samples across SDM runs in the unsupervised setting. Panels **a-d** show SDMs with increasing robustness to random seeds on CheXpertPlus (SDMs at {0, 33, 66, 100} percentiles). Rows and columns are ordered by the final consensus labels. Robust methods exhibit strong within-group co-occurrence (bright diagonal blocks) and low between-group co-occurrence (darker off-diagonal regions).

`K-Means` on CheXpertPlus, Fig. 3 **d**), the overall results underscore the need for consensus clustering in any practical subgroup discovery pipeline.

Domain adaptation demands differ from those of other tasks. SDMs that best combined target model guidance and external features performed best at domain adaptation (Tab. 1, "OOD" columns). On HAM10000 to BCN20000, the baseline OOD performance (without retraining) was 69.0 AUC, and the upper bound (when training on OOD directly) was 79.3 AUC. Here, `DOM-Bio(w10)` led with 70.2, followed by `K-Means` at 70.1. On CheXpert to BRAX, the baseline OOD performance was 82.2 AUC with an upper bound of 91.4 AUC (held-out BRAX split) with `STB` reaching 84.5 AUC, followed by `DOM(w10)`. Access to task labels did not translate to improvements in OOD performance.

3.3 Overall ranking: the simple baseline is surprisingly competitive

We aggregated pairwise comparisons across all tasks and datasets using Bradley-Terry ratings (Tab. 1, column "BT"). `K-Means` achieved the highest overall rating (2.71) in the unsupervised setting, followed by `DOM(w10)`. Under the BT model, `K-Means` had an average win probability of 85% against all other methods, and a 74% win probability against its closest competitor. Even more surprisingly, `K-Means` was also competitive in the supervised setting despite not having access to ground-truth labels and being much simpler to run than its SDM counterparts.

4 Discussion

Using our proposed framework for comprehensive evaluation of SDMs, we benchmarked five state-of-the-art methods on practical downstream tasks in medical imaging and assessed their robustness to random initialization. Our findings challenge conventional assumptions about subgroup discovery. No single method

dominated across all tasks, suggesting that different downstream applications impose different requirements on the discovered subgroup structure. When the target task is known in advance, practitioners may consult our per-task rankings to select an appropriate method. We find that simple `K-Means` is a surprisingly strong default, achieving the highest overall Bradley-Terry rating in unsupervised settings and the second-highest in supervised ones.

A recurring challenge we identify is the sensitivity of SDMs to random seeds. We propose a consensus clustering procedure that stabilizes subgroup assignments across runs. We recommend this as a default post-processing step when deploying SDMs in practice.

Our results show that SDMs provide meaningful benefits across all investigated tasks, demonstrating their utility beyond the typical setting of identifying spurious correlations, even without access to ground-truth target labels. Looking ahead, we anticipate that our comprehensive evaluation framework will accelerate the development of new SDMs by providing a clinically relevant benchmark. We plan to release a live leaderboard to support continued benchmarking as new methods and tasks emerge.

Acknowledgments. This project was supported by the Diabetes Center Berne (A.B., T.H., T.F., L.K.), strategic funding of the medical faculty of the University of Bern (T.H.), Natural Sciences and Engineering Council of Canada (NSERC) Postdoctoral Research Award (E.A.M.S.), the Royal Academy of Engineering as part of the Kheiron/RAEng Research Chair (B.G.). Calculations were performed on UBELIX, the HPC cluster at the University of Bern.

Disclosure of Interests. B.G. is a part-time employee of DeepHealth. No other competing interests.

References

1. Alloula, A., Jones, C., Glocker, B., Papiez, B.: Subgroups matter for robust bias mitigation. In: International Conference on Machine Learning (ICML) (2025)
2. Bissoto, A., Hoang, T.D., Flühmann, T., Sun, S., Baumgartner, C.F., Koch, L.M.: Subgroup performance analysis in hidden stratifications. In: International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI). pp. 594–603. Springer (2025)
3. Bradley, R.A., Terry, M.E.: Rank analysis of incomplete block designs: I. the method of paired comparisons. Biometrika **39**(3/4), 324–345 (1952)
4. Chambon, P., Delbrouck, J.B., Sounack, T., Huang, S.C., Chen, Z., Varma, M., Truong, S.Q., Chuong, C.T., Langlotz, C.P.: Chexpert plus: Augmenting a large chest x-ray dataset with text radiology reports, patient demographics and additional image formats. arXiv preprint arXiv:2405.19538 (2024)
5. Chiang, W.L., Zheng, L., Sheng, Y., Angelopoulos, A.N., Li, T., Li, D., Zhu, B., Zhang, H., Jordan, M., Gonzalez, J.E., et al.: Chatbot arena: An open platform for evaluating llms by human preference. In: International Conference on Machine Learning (ICML) (2024)

6. Combalia, M., Codella, N., Rotemberg, V., Helba, B., Vilaplana, V., Reiter, O., Halpern, A., Puig, S., Malvehy, J.: BCN20000: Dermoscopic lesions in the wild. arXiv:1908.02288 (2019)
7. Eyuboglu, S., Varma, M., Saab, K.K., Delbrouck, J.B., Lee-Messer, C., Dunnmon, J., Zou, J., Re, C.: Domino: Discovering systematic errors with cross-modal embeddings. In: International Conference on Learning Representations (2022)
8. Ghosh, S., Syed, R., Wang, C., Choudhary, V., Li, B., Poynton, C.B., Visweswaran, S., Batmanghelich, K.: Ladder: Language-driven slice discovery and error rectification in vision classifiers. In: Findings of the Association for Computational Linguistics (ACL Findings). pp. 22935–22970 (2025)
9. Guimard, Q., D'Incà, M., Mancini, M., Ricci, E.: Classifier-to-bias: Toward unsupervised automatic bias detection for visual classifiers. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 15151–15161 (2025)
10. Kurtansky, N.R., D'Alessandro, B.M., Gillis, M.C., Betz-Stablein, B., Cerminara, S.E., Garcia, R., Girundi, M.A., Goessinger, E.V., Gottfrois, P., Guitera, P., et al.: The slice-3d dataset: 400,000 skin lesion image crops extracted from 3d tbp for skin cancer detection. Scientific Data **11**(1), 884 (2024)
11. Mehta, R., Filos, A., Gal, Y., Arbel, T.: Uncertainty Evaluation Metric for Brain Tumour Segmentation. In: Medical Imaging with Deep Learning (MIDL) (2020)
12. Mehta, R., Shui, C., Arbel, T.: Evaluating the fairness of deep learning uncertainty estimates in medical image analysis. In: Medical Imaging with Deep Learning (MIDL) (2023)
13. Monti, S., Tamayo, P., Mesirov, J., Golub, T.: Consensus clustering: a resampling-based method for class discovery and visualization of gene expression microarray data. Machine learning **52**(1), 91–118 (2003)
14. Oakden-Rayner, L., Dunnmon, J., Carneiro, G., Re, C.: Hidden stratification causes clinically meaningful failures in machine learning for medical imaging. In: ACM Conference on Health, Inference, and Learning (2020)
15. Olesen, V., Weng, N., Feragen, A., Petersen, E.: Slicing through bias: explaining performance gaps in medical image analysis using slice discovery methods. In: MICCAI Workshop on Fairness of AI in Medical Imaging. pp. 3–13. Springer (2024)
16. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: International Conference on Machine Learning (ICML). pp. 8748–8763. PMLR (2021)
17. Reis, E.P., Paiva, J., Bueno da Silva, M.C., Sousa Ribeiro, G.A., Fornasiero Paiva, V., Bulgarelli, L., Lee, H., dos Santos, P.V., brito, v., Amaral, L., Beraldo, G., Haidar Filho, J.N., Teles, G., Szarf, G., Pollard, T., Johnson, A., Celi, L.A., Amaro, E.: BRAX, a Brazilian labeled chest X-ray dataset. PhysioNet (Jun 2022). https://doi.org/10.13026/grwk-yh18, https://doi.org/10.13026/grwk-yh18, version 1.1.0
18. Roschewitz, M., Mehta, R., Jones, C., Glocker, B.: Automatic dataset shift identification to support safe deployment of medical imaging ai. In: International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI). pp. 67–76. Springer (2025)
19. Stanley, E.A., Souza, R., Winder, A.J., Gulve, V., Amador, K., Wilms, M., Forkert, N.D.: Towards objective and systematic evaluation of bias in artificial intelligence for medical imaging. Journal of the American Medical Informatics Association **31**(11), 2613–2621 (2024)

20. Tschandl, P., Rosendahl, C., Kittler, H.: The ham10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions. Scientific data **5**(1) (2018)
21. Varma, M., Delbrouck, J.B., Chen, Z., Chaudhari, A., Langlotz, C.: Ravl: Discovering and mitigating spurious correlations in fine-tuned vision-language models. Advances in Neural Information Processing Systems (NeurIPS) (2024)
22. Weng, N., Petersen, E., Bissoto, A., Sun, S., Koch, L.M., Feragen, A., Bigdeli, S., Baumgartner, C.F.: Seg4seg: Identifying systematic failure modes in segmentation by subgroup discovery methods. In: Medical Imaging with Deep Learning (MIDL) (2026)
23. Wundram, A.M., Baumgartner, C.F.: Is uncertainty quantification a viable alternative to learned deferral? In: International Workshop on Uncertainty for Safe Utilization of Machine Learning in Medical Imaging (MICCAI UNSURE Workshop). pp. 34–44. Springer (2025)
24. Yenamandra, S., Ramesh, P., Prabhu, V., Hoffman, J.: Facts: First amplify correlations and then slice to discover bias. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 4794–4804 (2023)
25. Zhang, S., Xu, Y., Usuyama, N., Xu, H., Bagga, J., Tinn, R., Preston, S., Rao, R., Wei, M., Valluri, N., Wong, C., Tupini, A., Wang, Y., Mazzola, M., Shukla, S., Liden, L., Gao, J., Crabtree, A., Piening, B., Bifulco, C., Lungren, M.P., Naumann, T., Wang, S., Poon, H.: A multimodal biomedical foundation model trained from fifteen million image–text pairs. NEJM AI **2**(1) (2024). https://doi.org/10.1056/AIoa2400640, https://ai.nejm.org/doi/full/10.1056/AIoa2400640