Clinical Uncertainty Impacts Machine Learning Evaluations

Simone Lionetti[*1] SIMONE.LIONETTI@HSLU.CH
Fabian Gröger[*2,1] FABIAN.GROEGER@UNIBAS.CH
Philippe Gottfrois[2] PHILIPPE.GOTTFROIS@UNIBAS.CH
Alvaro Gonzalez-Jimenez[3] ALVARO.GONZALEZJIMENEZ@UNIBAS.CH
Ludovic Amruthalingam[1] LUDOVIC.AMRUTHALINGAM@HSLU.CH
Alexander A. Navarini[†2,3] ALEXANDER.NAVARINI@USB.CH
Marc Pouly[†1] MARC.POULY@HSLU.CH
[1] *Department of Computer Science, Lucerne University of Applied Sciences and Arts, Switzerland*
[2] *Department of Biomedical Engineering, University of Basel, Switzerland*
[3] *Department of Dermatology, University Hospital of Basel, Switzerland*

Abstract

Clinical dataset labels are rarely certain as annotators disagree and confidence is not uniform across cases. Typical aggregation procedures, such as majority voting, obscure this variability. In simple experiments on medical imaging benchmarks, accounting for the confidence in binary labels significantly impacts model rankings. We therefore argue that machine-learning evaluations should explicitly account for annotation uncertainty using probabilistic metrics that directly operate on distributions. These metrics can be applied independently of the annotations' generating process, whether modeled by simple counting, subjective confidence ratings, or probabilistic response models. They are also computationally lightweight, as closed-form expressions have linear-time implementations once examples are sorted by model score. We thus urge the community to release raw annotations for datasets and to adopt uncertainty-aware evaluation so that performance estimates may better reflect clinical data.

Keywords: Evaluation, annotation, uncertainty, metrics.

Data and Code Availability. Datasets are publicly accessible and detailed in Section 2.

Institutional Review Board (IRB). IRB approval is not required for our work, as it utilizes existing datasets and annotations without further interaction with human subjects.

1. Introduction

The availability of health-related datasets has grown rapidly (Kiryati and Landau, 2021). These resources have been instrumental in advancing machine learning (ML) for healthcare, enabling the community to benchmark methods and accelerate progress (Johnson et al., 2016; Tschandl et al., 2018; Irvin et al., 2019). A key barrier to their utility is the uncertainty intrinsic to clinical annotations. Specifically, even among domain experts, agreement on the presence or absence of a finding is often low, reflecting the ambiguity of medical data and the subjectivity of interpretation (Elmore et al., 2015; Krause et al., 2018). To mitigate this, it has become common practice to collect multiple annotations per sample (Armato III et al., 2011; Irvin et al., 2019; Raumanns et al., 2021).

For example in dermatology even histopathology, treated as gold standard diagnosis, achieves only moderate agreement. Specifically, in an observational study of 60 melanoma cases across three Spanish hospitals, mean inter-observer agreement gave a Cohen's κ around 0.5 (Sanz-Motilva et al., 2025). This variability reflects intrinsic ambiguity. Forcing deterministic labels obscures it and biases evaluation against uncertainty-aware models.

While multi-annotator designs acknowledge the uncertainty of the labeling process, the resulting annotations are typically aggregated into a single "ground-truth" label, often by majority voting or thresholding (Snow et al., 2008). This produces an illusion of certainty: An image labeled as "positive" by 6/10 experts is treated identically to one unanimously labeled as "positive" by 10/10. Even in the case of 2-5

[*] Joint first authorship.
[†] Joint last authorship.

annotators, which is more common in health where expert annotations are costly, a 2/3 agreement conveys critically different information than a unanimous 3/3 agreement. This nuance is lost with majority voting or thresholding. Current ML pipelines often evaluate models against aggregated labels as if they were certain, disregarding the underlying uncertainty (Irvin et al., 2019; Chen et al., 2021). In doing so, meaningful variations in expert opinion are collapsed into a binary outcome, obscuring the fact that evaluation is performed against a fragile construct rather than a real reference.

We argue that this practice is not aligned with the nuances of clinical data. Evaluation should incorporate uncertainty, as ignoring it leads to a hidden selection bias, where models that more closely align with thresholded labels are favored over those that predict realistic uncertainty. Importantly, doing so is neither conceptually nor technically difficult, and has a sizeable impact on results. There exist extensions of widely used metrics, such as area under the receiver operating characteristic curve (AUROC) or average precision (AP), that apply to probabilistic labels. These can be traced back at least 20 years in the information retrieval literature (Kekäläinen and Järvelin, 2002), but remain rarely applied in the ML for health community.

Rather than collapsing disagreement to a binary label, uncertainty-aware *soft* metrics directly operate on continuous probabilities in the $[0, 1]$ range. Two properties make them immediately practical. First, soft metrics are agnostic to the assumptions used to obtain probability estimates, which can range from independent votes or subjective confidence to item-response theory. Second, they are computationally tractable, as closed-form expressions allow for linear-time execution after sorting by score, just as in the binary case.

Related work. The issue of annotation uncertainty and its impact on ML model evaluation, especially for clinical tasks, has been highlighted several times. Maier-Hein et al. (2018) showed that rankings in biomedical image analysis challenges fluctuate with annotator selection and aggregation scheme, urging for transparency with respect to label uncertainty. For clinical applications, Chen et al. (2021) argued that when the reference standard is subjective, agreement should be measured with human comparators, avoiding claims of accuracy against unquestioned truth. Gordon et al. (2021) took the program further by addressing intra-annotator variations and then averaging metrics across annotators. A community perspective (Reinke et al., 2024) emphasized that metric choice and aggregation must align with the problem and data, highlighting pitfalls that arise when subjectivity is ignored.

Several thresholdless metrics have been proposed to address uncertain annotations. The information retrieval formulation of precision and recall to continuous, non-binary labels included the soft version of AP (Kekäläinen and Järvelin, 2002). A similar approach was used to evaluate boundary detection in images by taking the average of scores over different annotators (Martin et al., 2004). Further works formulated precision and recall with frequencies as probabilities, including all-negative and all-positive dummies to avoid certainty when labels are classifier outputs, and investigated hypothesis testing on these metrics (Lamiroy and Sun, 2011; Lamiroy and Pierrot, 2015).

Contributions. This paper draws attention to the gap between annotation practice and evaluation methodology. Through evaluations on benchmark medical imaging datasets, we show that accounting for label uncertainty can substantially alter the ranking of competing methods. This evidence underscores the need for a shift away from evaluation with artificially certain labels towards faithful representation of annotation uncertainty. We compile explicit, simple expressions to compute soft versions of AP and AUROC that are cheap to compute and straightforward to interpret, thus facilitating their adoption. We conclude by urging the community to follow uncertainty-aware evaluation practices and to promote transparency by releasing unaggregated annotations.

2. Experiments

2.1. Experimental setup

Metrics. Let items $1, \dots, n$ be sorted in descending order by their score for positive classification. Denote the probability of item i being positive by $p_i \in [0, 1]$. Define the cumulative counts

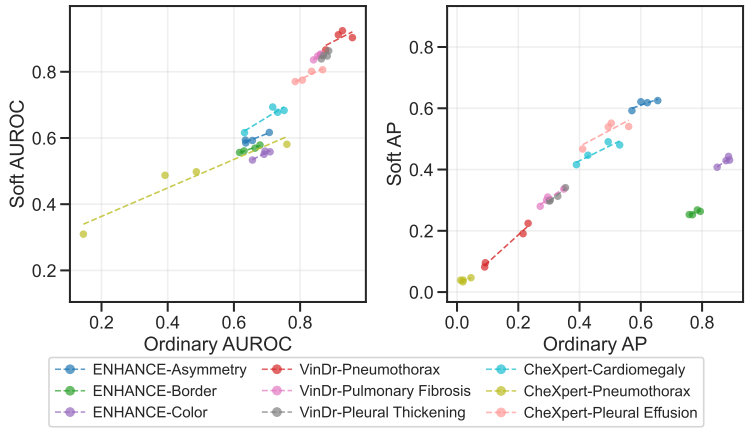
$$n_i^+ = \sum_{j=1}^{i} p_j, \qquad n_i^- = \sum_{j=1}^{i} (1 - p_j), \tag{1}$$

and totals $n^+ = n_n^+$, $n^- = n_n^-$. True positive rate (TPR) and false positive rate (FPR) are defined by

$$\mathrm{TPR}_i = n_i^+ / n^+, \qquad \mathrm{FPR}_i = n_i^- / n^-, \tag{2}$$

and precision and recall by

$$\mathrm{P}_i = n_i^+ / i, \qquad \mathrm{R}_i = n_i^+ / n^+. \tag{3}$$



Dataset	R^2 AUROC	R^2 AP
ENHANCE-Asymmetry	0.899	0.697
ENHANCE-Border	0.966	0.695
ENHANCE-Color	0.874	0.834
VinDr-Pleural Thickening	0.697	0.969
VinDr-Pneumothorax	0.465	0.983
VinDr-Pulmonary Fibrosis	0.982	0.941
CheXpert-Cardiomegaly	0.840	0.856
CheXpert-Pleural Effusion	0.897	0.726
CheXpert-Pneumothorax	0.916	0.659

Figure 1: Comparison of ordinary and soft metrics on three datasets with three tasks each. In the left panels, dots represent different backbones, and dashed lines indicate Pearson score correlations whose R^2 values are reported on the right. Details are in Appendix Table 2.

Definition 1 (Soft AUROC)
The soft (uncertainty-aware) AUROC is defined as

$$\begin{aligned} \text{s-AUROC} &= \sum_{i=1}^{n} \text{TPR}_i(\text{FPR}_i - \text{FPR}_{i-1}) \\ &= \sum_{i=1}^{n} \sum_{j=1}^{i-1} \frac{(1-p_i)p_j}{n^+ n^-}. \end{aligned} \tag{4}$$

Definition 2 (Soft AP)
The soft (uncertainty-aware) AP is defined as

$$\text{s-AP} = \sum_{i=1}^{n} \text{P}_i(\text{R}_i - \text{R}_{i-1}) = \sum_{i=1}^{n} \sum_{j=1}^{i} \frac{p_i}{i} \frac{p_j}{n^+}. \tag{5}$$

These definitions automatically reduce to ordinary AUROC and AP when labels are binary. Using the cumulative sums in the first lines of the equations allows for linear time implementations, assuming the samples are already sorted by score. The metric values are deterministic and do not require Monte Carlo sampling, despite taking into account the probabilistic nature of the annotation process.

In the experiments that follow, we compare *ordinary* AUROC and AP computed against *binary* labels, and their *soft* counterparts s-AUROC and s-AP computed directly from *probabilistic* labels.

Tasks and datasets. We first evaluate soft ranking metrics across three medical imaging settings with varying label uncertainty (see Appendix B). The first is the dermatology benchmark *ENHANCE* (Raumanns et al., 2021), featuring annotations of the student groups 2017–2020 for the lesion attributes asymmetry, border, and color. For each image, we normalize ratings to $[0, 1]$ and average them to obtain a soft label, and binarize labels for presence of the attribute by requiring a mean >1 on the original scale. We then consider the two chest X-ray benchmarks *VinDR* (Nguyen et al., 2022) and *CheXpert*[1] (Irvin et al., 2019). Here, we take the mean of expert annotations for each image to obtain soft targets and binarize them with majority voting, which aligns with prior work on this dataset (Irvin et al., 2019). For datasets without a predefined test split, we select a 30% test set with a fixed random seed.

To evaluate how uncertainty affects evaluation across a broader range of tasks beyond typical medical imaging, and beyond simple averaging for probabilistic labels, we also investigate data quality issue detection on *CleanPatrick* (Gröger et al., 2025). This benchmark features raw, unaggregated annotations for three issue types on medical images (off-topic samples, near duplicates, and label errors). We follow the same setup and evaluation as the benchmark, including the evaluated methods and GLAD aggregation (see Appendix E).

Features and models. We extract image representations with four backbones pretrained on ImageNet with supervision: VGG-16 (Simonyan and Zisserman, 2015), ResNet-50 (He et al., 2016), EfficientNet-b0 (Tan and Le, 2019), and ViT-base

1. Only the test set contains multi-rater annotations.

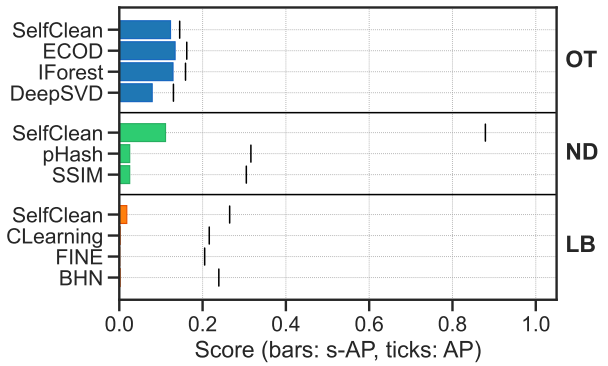


Figure 2: Results for data quality issue detection on *CleanPatrick*. Bars show the uncertainty-aware score s-AP, and vertical ticks mark the corresponding AP. Detailed results are in Table 2 of the appendix.

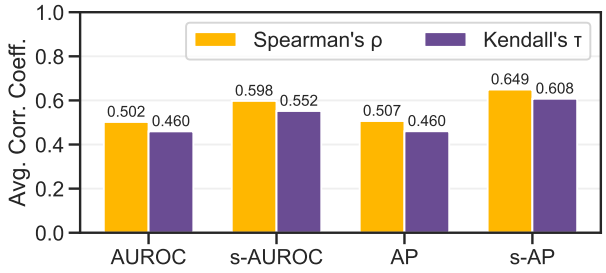


Figure 3: Average correlation coefficient of rankings produced by ordinary and soft metrics upon annotation bootstrap across 13 tasks. Details are in Figure 5 of the appendix.

(Dosovitskiy et al., 2021). Following the methods' preprocessing, all images are resized to 224 × 224, converted to tensors, and normalized with ImageNet mean and standard deviation. *VinDr*'s DICOM files are read using `pydicom` (Mason and the pydicom contributors, 2024). Pixel arrays are scaled to 8-bit if needed, and single-channel images are repeated to RGB. We train a standard Logistic Regression classifier from scikit-learn on the features extracted with frozen backbones and report probabilities on the held-out test set. While more sophisticated and performant backbones exist, these models remain relevant in practice for clinical settings, where computational infrastructure may be limited.

2.2. Results

Switching from ordinary to uncertainty-aware evaluation reweights ambiguous cases. Besides decreasing absolute scores and compressing their range, this importantly reshuffles leaders across datasets as can be seen in Figure 1. On *ENHANCE* Border/Color, ordinary APs above 0.75 collapse to lower and tighter s-AP regions. In this regime, VGG-16 outperforms other backbones, indicating that the best method does not merely separate binary positives, but rather correctly handles borderline lesions. A similar pattern can be observed on *VinDr*, where soft metrics are uniformly lower yet still flip winners on several tasks (*e.g.*, VGG-16 performs best on Pneumothorax, Pulmonary Fibrosis, and Pleural Thickening), underscoring that rank is sensitive to how label uncertainty is modeled. The analysis of the coefficients of determination R^2 in Figure 1 shows that AP often decouples rankings between ordinary and soft settings moderately, and AUROC can sometimes decouple them sharply (*e.g.*, CheXpert–Pneumothorax AUROC $R^2 = 0.465$ vs. AP $R^2 = 0.983$).

Note that most top-ranked solutions in the *VinDr* challenge ensemble multiple models, often including pretrained and fine-tuned ResNets. Even with frozen backbones and a linear head, VGG-16 achieves better results than ResNet counterparts when evaluated with uncertainty-aware metrics on several tasks. This suggests that its representations may lead to better logit calibration for evaluations that explicitly model annotation uncertainty, compared to the widely used fine-tuned ResNet variants.

On *CleanPatrick* (see Figure 2), uncertainty-aware evaluation does not change the best method but reorders competing ones across tasks. For *near duplicates*, SelfClean remains first, while pHash and SSIM collapse to a tie under s-AP. For *off-topic samples*, the ordering is mostly unchanged. The starkest shift is observed for *label errors*: BHN sits second in terms of ordinary AP but drops near the bottom under s-AP, while SelfClean keeps first rank. Overall, shifting from ordinary to soft metrics changes which models are better than others in ambiguous cases.

In Figure 3, we further investigate the reliability of the ordinary and soft metrics. We perform a bootstrap analysis as discussed in Appendix C and compare the stability of the model rankings by measuring the correlation coefficient among them. The results show that using soft metrics leads to more consistent evaluation across all tasks.

To ensure that findings generalize beyond linear evaluation on frozen features, we also conduct experiments on Chexpert both with end-to-end fine-tuning (Table 3) of the same model setup (i.e., backbone and linear head) and using leaderboard models (Table 4). These results, detailed in Appendix D, confirm that uncertainty-aware metrics frequently reshuffle model rankings including the top position.

3. Discussion

This work considers uncertainty in ranking metrics for binary classification. While this also applies to binary multi-label problems, extension to the multi-class case is only straightforward in a one-vs-one or a one-vs-all fashion.

Soft metrics are agnostic both to the modality and to the origin of probabilistic labels. They are directly applicable to any domain whenever probabilistic labels are available. Beyond the example of melanoma diagnosis (Elmore et al., 2017), other clinical domains show high inter-rater variability where soft metrics are advantageous. In radiology, the interpretation of ambiguous lung nodules or characterization of lesions in mammography often results in significant disagreement (Lehman et al., 2015). In pathology, the grading of tumors, like Gleason scoring for prostate cancer, is notoriously subjective (Egevad et al., 2013). In ophthalmology, severity grading of diabetic retinopathy leads to expert disagreement (Krause et al., 2018). These can be contrasted with domains where disagreement is typically lower, such as bone fracture detection (Nowroozi et al., 2024), to identify where uncertainty-aware evaluation is most critical.

The choice of aggregation model (e.g., simple averaging or item-response theory) may have a significant influence. These nuances are not fully explored yet, due to scoping and the limited availability of public medical datasets with unaggregated annotations. The work by Stutz et al. (2023) offers a complementary perspective to soft ranking metrics, as it describes how to produce prediction sets with confidence guarantees using conformal prediction.

One might also hypothesize that uncertainty-aware training will produce better results in terms of soft metrics, but evidence for this is yet to be collected. The surprising performance of VGG-16 in some tasks warrants discussion. One hypothesis is that this architecture may be less prone to overfitting binarized majority labels in mid-small datasets compared to more complex ones. Although monotonic calibration such as Platt scaling does not change the result of ranking metrics, models that are not overly confident and incorporate better uncertainty estimates are rewarded by disagreement-aware evaluation.

The full clinical impact of soft-metrics remains to be assessed. Results suggest that decisions about which models should be deployed are likely to change due to rank reshuffling. While incorporating uncertainty within evaluations is intuitive, downstream consequences should be investigated.

Finally, releasing unaggregated annotations may in some cases present challenges related to privacy, ethics, or legality, as discussed in Appendix A. On the other hand, it has great potential to improve fairness, for instance, by investigating if high-disagreement cohorts correlate with demographic factors.

4. Conclusion

Label uncertainty is intrinsic to clinical data. Uncertainty-aware ranking metrics that operate on probabilistic targets are easy to compute, improve the stability of rankings, and often change which models perform best.

Two practical recommendations follow. First, benchmark creators should release unaggregated annotations or at least fractional targets to enable uncertainty-aware evaluation. Second, practitioners should report uncertainty-aware metrics alongside ordinary ones and comment on any rank changes. These steps make empirical claims more robust to the irreducible ambiguity of clinical annotation.

Acknowledgments

This research was funded in part by the Swiss National Science Foundation (SNSF) under grant 20HW-1_228541.

References

Samuel G Armato III, Geoffrey McLennan, Luc Bidaut, Michael F McNitt-Gray, Charles R Meyer, Anthony P Reeves, Binsheng Zhao, Denise R Aberle, Claudia I Henschke, Eric A Hoffman, et al. The lung image database consortium (LIDC) and image database resource initiative (IDRI): a completed reference database of lung nodules on CT scans. *Medical Physics*, 2011.

Po-Hsuan Cameron Chen, Craig H Mermel, and Yun Liu. Evaluation of artificial intelligence on a reference standard based on subjective interpretation. *The Lancet Digital Health*, 2021.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. In *International Conference on Learning Representations*, 2021.

Lars Egevad, Amar S Ahmad, Ferran Algaba, Daniel M Berney, Liliane Boccon-Gibod, Eva Compérat, Andrew J Evans, David Griffiths, Rainer Grobholz, Glen Kristiansen, et al. Standardization of gleason grading among 337 european pathologists. *Histopathology*, 2013.

Joann G Elmore, Gary M Longton, Patricia A Carney, Berta M Geller, Tracy Onega, Anna NA Tosteson, Heidi D Nelson, Margaret S Pepe, Kimberly H Allison, Stuart J Schnitt, et al. Diagnostic concordance among pathologists interpreting breast biopsy specimens. *JAMA*, 2015.

Joann G Elmore, Raymond L Barnhill, David E Elder, Gary M Longton, Margaret S Pepe, Lisa M Reisch, Patricia A Carney, Linda J Titus, Heidi D Nelson, Tracy Onega, et al. Pathologists' diagnosis of invasive melanoma and melanocytic proliferations: observer accuracy and reproducibility study. *bmj*, 2017.

Mitchell L Gordon, Kaitlyn Zhou, Kayur Patel, Tatsunori Hashimoto, and Michael S Bernstein. The disagreement deconvolution: Bringing machine learning performance metrics in line with reality. In *CHI Conference on Human Factors in Computing Systems*, 2021.

Fabian Gröger, Simone Lionetti, Philippe Gottfrois, Alvaro Gonzalez-Jimenez, Ludovic Amruthalingam, Labelling Consortium, Matthew Groh, Alexander A. Navarini, and Marc Pouly. Intrinsic Self-Supervision for Data Quality Audits. *Advances in Neural Information Processing Systems*, 2024.

Fabian Gröger, Simone Lionetti, Philippe Gottfrois, Alvaro Gonzalez-Jimenez, Ludovic Amruthalingam, Elisabeth Victoria Goessinger, Hanna Lindemann, Marie Bargiela, Marie Hofbauer, Omar Badri, et al. CleanPatrick: A Benchmark for Image Data Cleaning. *arXiv*, 2025.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2016.

Jeremy Irvin, Pranav Rajpurkar, Michael Ko, Yifan Yu, Silviana Ciurea-Ilcus, Chris Chute, Henrik Marklund, Behzad Haghgoo, Robyn Ball, Katie Shpanskaya, et al. Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison. In *AAAI Conference on Artificial Intelligence*, 2019.

Alistair EW Johnson, Tom J Pollard, Lu Shen, Li-wei H Lehman, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G Mark. MIMIC-III, a freely accessible critical care database. *Scientific Data*, 2016.

Jaana Kekäläinen and Kalervo Järvelin. Using graded relevance assessments in IR evaluation. *Journal of the American Society for Information Science and Technology*, 2002.

Taehyeon Kim, Jongwoo Ko, JinHwan Choi, Se-Young Yun, et al. Fine samples for learning with noisy labels. *Advances in Neural Information Processing Systems*, 2021.

Nahum Kiryati and Yuval Landau. Dataset growth in medical image analysis research. *Journal of Imaging*, 2021.

Jonathan Krause, Varun Gulshan, Ehsan Rahimy, Peter Karth, Kasumi Widner, Greg S Corrado, Lily Peng, and Dale R Webster. Grader variability and the importance of reference standards for evaluating machine learning models for diabetic retinopathy. *Ophthalmology*, 2018.

Bart Lamiroy and Pascal Pierrot. Statistical performance metrics for use with imprecise ground-truth. In *International Workshop on Graphics Recognition*, 2015.

Bart Lamiroy and Tao Sun. Computing precision and recall with missing or uncertain ground truth. In *International Workshop on Graphics Recognition*, 2011.

Constance D Lehman, Robert D Wellman, Diana SM Buist, Karla Kerlikowske, Anna NA Tosteson, Diana L Miglioretti, Breast Cancer Surveillance Consortium, et al. Diagnostic accuracy of digital screening mammography with and without computer-aided detection. *JAMA internal medicine*, 2015.

Zheng Li, Yue Zhao, Xiyang Hu, Nicola Botta, Cezar Ionescu, et al. ECOD: Unsupervised Outlier Detection Using Empirical Cumulative Distribution Functions. *IEEE Transactions on Knowledge and Data Engineering*, 2022.

Fei Tony Liu, Kai Ming Ting, and Zhi-Hua Zhou. Isolation Forest. In *IEEE International Conference on Data Mining*, 2008.

Lena Maier-Hein, Matthias Eisenmann, Annika Reinke, Sinan Onogur, Marko Stankovic, Patrick Scholz, Tal Arbel, Hrvoje Bogunovic, Andrew P Bradley, Aaron Carass, et al. Why rankings of biomedical image analysis competitions should be interpreted with care. *Nature Communications*, 2018.

David R Martin, Charless C Fowlkes, and Jitendra Malik. Learning to detect natural image boundaries using local brightness, color, and texture cues. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2004.

Darcy L. Mason and the pydicom contributors. pydicom v3.0.1, 2024.

Ha Q Nguyen, Khanh Lam, Linh T Le, Hieu H Pham, Dat Q Tran, Dung B Nguyen, Dung D Le, Chi M Pham, Hang TT Tong, Diep H Dinh, et al. Vindr-cxr: An open dataset of chest x-rays with radiologist's annotations. *Scientific Data*, 2022.

Curtis Northcutt, Lu Jiang, and Isaac Chuang. Confident learning: Estimating uncertainty in dataset labels. *Journal of Artificial Intelligence Research*, 2021.

A Nowroozi, MA Salehi, P Shobeiri, S Agahi, S Momtazmanesh, P Kaviani, and MK Kalra. Artificial intelligence diagnostic accuracy in fracture detection from plain radiographs and comparing it with clinicians: a systematic review and meta-analysis. *Clinical Radiology*, 2024.

Ralf Raumanns, Gerard Schouten, Max Joosten, Josien PW Pluim, Veronika Cheplygina, et al. ENHANCE (ENriching Health data by ANnotations of Crowd and Experts): A case study for skin lesion classification. *Machine Learning for Biomedical Imaging*, 2021.

Annika Reinke, Minu D Tizabi, Michael Baumgartner, Matthias Eisenmann, Doreen Heckmann-Nötzel, A Emre Kavur, Tim Rädsch, Carole H Sudre, Laura Acion, Michela Antonelli, et al. Understanding metric-related pitfalls in image analysis validation. *Nature Methods*, 2024.

Lukas Ruff, Robert Vandermeulen, Nico Goernitz, Lucas Deecke, Shoaib Ahmed Siddiqui, Alexander Binder, Emmanuel Müller, and Marius Kloft. Deep one-class classification. In *International Conference on Machine Learning*, 2018.

V Sanz-Motilva, A Martorell, E Manrique-Silva, L Terradez-Mas, C Requena, V Traves, O Sanmartín, JL Rodríguez-Peralto, and E Nagore. Interobserver Variability in the Histopathological Evaluation of Melanoma: Analysis of 60 Cases. *Actas Dermo-Sifiliográficas*, 2025.

Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. In *International Conference on Learning Representations*, 2015.

Rion Snow, Brendan O'connor, Dan Jurafsky, and Andrew Y Ng. Cheap and fast–but is it good? evaluating non-expert annotations for natural language tasks. In *Conference on Empirical Methods in Natural Language Processing*, 2008.

David Stutz, Abhijit Guha Roy, Tatiana Matejovicova, Patricia Strachan, Ali Taylan Cemgil, and Arnaud Doucet. Conformal prediction under ambiguous ground truth. *Transactions on Machine Learning Research*, 2023.

Mingxing Tan and Quoc V. Le. Rethinking model scaling for convolutional neural networks. In *International Conference on Machine Learning*, 2019.

Philipp Tschandl, Cliff Rosendahl, and Harald Kittler. The HAM10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions. *Scientific Data*, 2018.

Zhou Wang, A.C. Bovik, H.R. Sheikh, and E.P. Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE Transactions on Image Processing*, 2004.

Chenglin Yu, Xinsong Ma, and Weiwei Liu. Delving into noisy label detection with clean data. In *International Conference on Machine Learning*, 2023.

Christoph Zauner. Implementation and benchmarking of perceptual image hash functions. 2010.

Appendix A. Barriers to data sharing

In many practical cases, releasing individual annotations with anonymized annotator identifiers entails low ethical, legal, or privacy risks. Annotators are usually professionals participating with informed consent rather than vulnerable subjects, and the additional metadata does not expose new health information beyond what is already available. When identifiers are randomized and links to institutions or timestamps are loose, re-identification of either patients or annotators requires substantial prior knowledge and effort. Certain datasets may warrant additional precautions due to specific contextual factors, but this is not the norm. Provided the release agreement forbids misuse and a standard data-protection assessment confirms anonymization, the potential for harm remains remote, whereas the gains for clinical research are substantial.

Appendix B. Uncertainty distributions

Figure 4 shows the distributions of soft and hard labels for the evaluated datasets of the main paper. We can see that *ENHANCE* has the highest uncertainty, which can be attributed to the large amounts of annotations per sample. In this case, collapsing the labels to a binary outcome results in the loss of a significant amount of information. *CheXpert* contains eight annotations per sample, while for *VinDr* we found three annotations per sample, resulting in less uncertainty.

Appendix C. Bootstrap analysis details

For the bootstrap analysis, we resample the annotations for each image with replacement to generate a new set of labels for the entire dataset. We then recalculate all metrics (AUROC, AP, s-AUROC, s-AP) for each model, re-rank the models based on the new scores, and measure the rank correlation with the original order using Spearman's ρ and Kendall's τ. We repeat this 1,000 times and use the average rank correlation as an estimate of stability.

Table 1: Binomial test p-values comparing the stability (average rank correlation) of ordinary vs. soft metrics over 1,000 bootstrap iterations. Values < 0.05 indicate the soft metric is significantly more stable. We see this in 8/9 tasks for AP and 3/9 for AUROC.

Dataset	AP vs. s-AP	AUROC vs. s-AUROC
CheXpert-Pleural Effusion	0.0119	0.2135
CheXpert-Pneumothorax	>0.9999	>0.9999
VinDr-Pleural Thickening	<0.0001	0.8019
CheXpert-Cardiomegaly	<0.0001	0.2141
VinDr-Pulmonary Fibrosis	<0.0001	<0.0001
VinBigData-Pneumothorax	0.0001	0.0138
ENHANCE-Color	0.0036	0.2971
ENHANCE-Border	<0.0001	0.0049
ENHANCE-Asymmetry	<0.0001	0.0625

Appendix D. Extended results

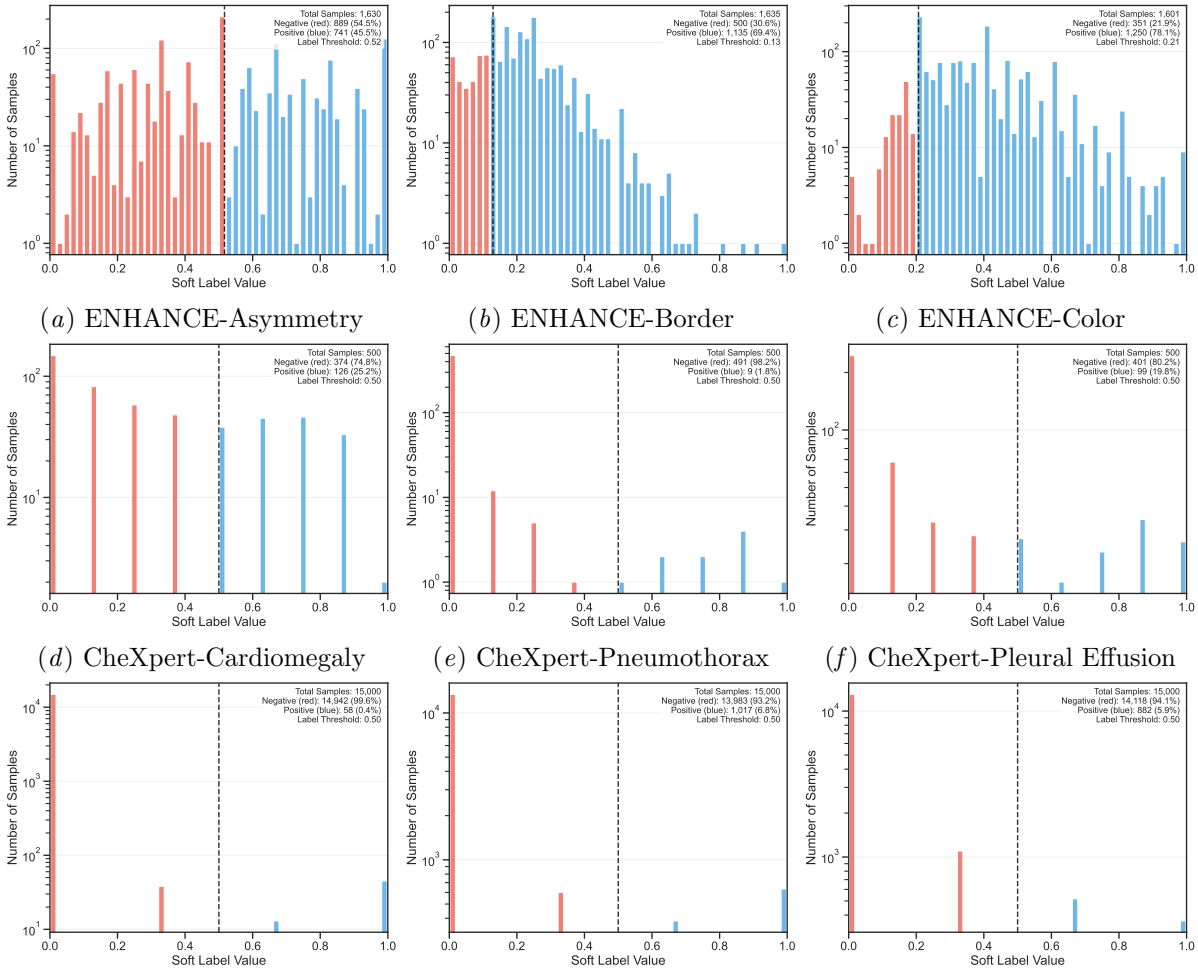
We provide additional results on the evaluations performed in the main paper.

Table 2 presents the performance evaluation in terms of ordinary and soft metrics for various datasets, specifically for the three attribute detection tasks on *ENHANCE*, five disease prediction tasks on *VinDr*, five disease prediction tasks on *CheXpert*, and the three data quality issue detection tasks on *CleanPatrick*. To ensure these findings generalize, Table 3 shows the results of end-to-end fine-tuning, and Table 4 details the re-ranking of external leaderboard models from CheXternal. Figure 6 visualizes the evaluation of soft versus ordinary metrics for linear probing, fine-tuning, and leaderboard models.

Figure 5 shows the results of the bootstrap analysis broken down by dataset, which expands on the aggregated Figure 3.

Appendix E. Description of cleaning approaches

We briefly outline the methods we evaluated to detect off-topic samples, near duplicates, and label errors. For implementation details, please refer to the cited



(a) ENHANCE-Asymmetry (b) ENHANCE-Border (c) ENHANCE-Color

(d) CheXpert-Cardiomegaly (e) CheXpert-Pneumothorax (f) CheXpert-Pleural Effusion

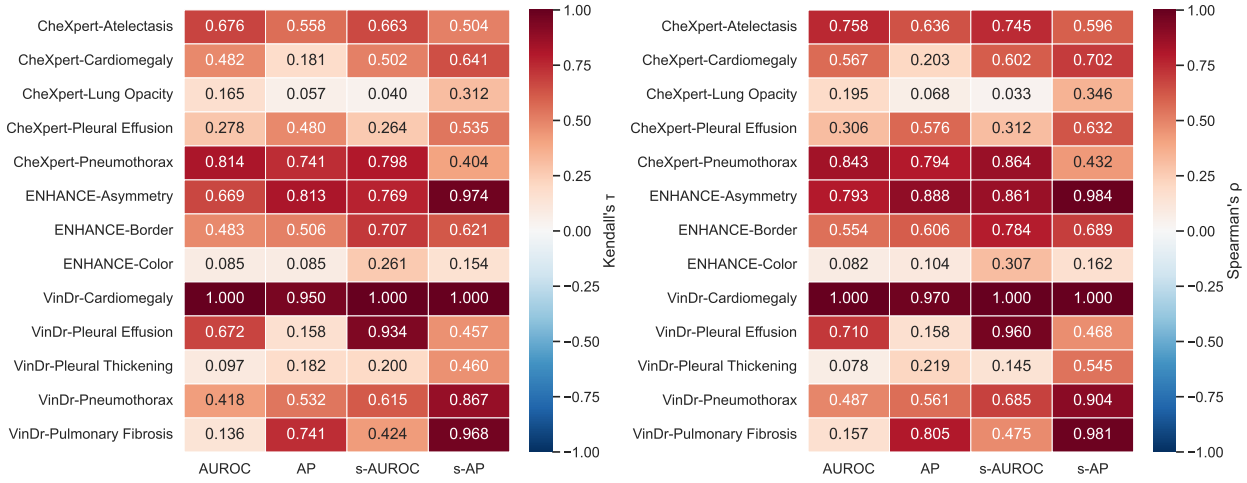
(g) VinDr-Pneumothorax (h) VinDr-Pulmonary Fibrosis (i) VinDr-Pleural Thickening

Figure 4: Distribution of soft labels for the evaluated datasets. Dotted lines represent the threshold used to produce the hard labels, and the colors represent binary positive and negative labels.

Table 2: Comparison of performance measured in terms of ordinary vs. soft metrics for different methods and on various datasets.

	Ordinary Metrics		Soft Metrics	
Method	AUROC	AP	s-AUROC	s-AP
ENHANCE-Asymmetry				
EfficientNet-b0	0.636	0.571	0.585	0.592
ResNet-50	0.656	0.621	0.593	0.618
VGG-16	0.707	0.655	0.617	0.625
ViT-base	0.635	0.600	0.594	0.621
ENHANCE-Border				
EfficientNet-b0	0.664	0.794	0.569	0.263
ResNet-50	0.617	0.758	0.556	0.253
VGG-16	0.679	0.785	0.579	0.268
ViT-base	0.631	0.769	0.561	0.253
ENHANCE-Color				
EfficientNet-b0	0.709	0.890	0.558	0.431
ResNet-50	0.691	0.878	0.551	0.429
VGG-16	0.694	0.886	0.560	0.443
ViT-base	0.656	0.849	0.533	0.408
VinDr-Cardiomegaly				
EfficientNet-b0	0.935	0.672	0.931	0.653
ResNet-50	0.942	0.711	0.938	0.692
VGG-16	0.951	0.740	0.947	0.716
ViT-base	0.924	0.654	0.922	0.637
VinDr-Pneumothorax				
EfficientNet-b0	0.958	0.216	0.903	0.191
ResNet-50	0.877	0.090	0.866	0.082
VGG-16	0.928	0.232	0.924	0.224
ViT-base	0.915	0.093	0.912	0.096
VinDr-Pleural Effusion				
EfficientNet-b0	0.896	0.477	0.884	0.452
ResNet-50	0.921	0.494	0.905	0.468
VGG-16	0.926	0.564	0.910	0.529
ViT-base	0.914	0.495	0.894	0.457
VinDr-Pulmonary Fibrosis				
EfficientNet-b0	0.841	0.292	0.836	0.300
ResNet-50	0.853	0.272	0.848	0.280
VGG-16	0.862	0.349	0.854	0.336
ViT-base	0.861	0.296	0.851	0.310
VinDr-Pleural Thickening				
EfficientNet-b0	0.864	0.304	0.839	0.300
ResNet-50	0.870	0.302	0.849	0.297
VGG-16	0.886	0.354	0.863	0.341
ViT-base	0.882	0.329	0.848	0.313

	Ordinary Metrics		Soft Metrics	
Method	AUROC	AP	s-AUROC	s-AP
CheXpert-Cardiomegaly				
EfficientNet-b0	0.752	0.531	0.683	0.480
ResNet-50	0.732	0.427	0.678	0.447
VGG-16	0.717	0.493	0.694	0.491
ViT-base	0.632	0.390	0.616	0.416
CheXpert-Pneumothorax				
EfficientNet-b0	0.145	0.012	0.310	0.039
ResNet-50	0.392	0.020	0.487	0.039
VGG-16	0.486	0.019	0.498	0.033
ViT-base	0.760	0.046	0.581	0.047
CheXpert-Pleural Effusion				
EfficientNet-b0	0.835	0.504	0.801	0.551
ResNet-50	0.868	0.560	0.806	0.540
VGG-16	0.807	0.494	0.775	0.539
ViT-base	0.785	0.410	0.770	0.467
CheXpert-Lung Opacity				
EfficientNet-b0	0.847	0.828	0.813	0.813
ResNet-50	0.832	0.794	0.798	0.782
VGG-16	0.842	0.813	0.804	0.804
ViT-base	0.820	0.799	0.799	0.796
CheXpert-Atelectasis				
EfficientNet-b0	0.712	0.457	0.718	0.547
ResNet-50	0.765	0.515	0.729	0.555
VGG-16	0.742	0.504	0.735	0.596
ViT-base	0.686	0.389	0.679	0.496
CleanPatrick: Off-topic Samples				
SelfClean	0.669	0.145	0.676	0.123
ECOD	0.757	0.162	0.754	0.134
IForest	0.773	0.159	0.774	0.129
DeepSVD	0.728	0.130	0.714	0.079
CleanPatrick: Near Duplicates				
SelfClean	0.917	0.879	0.888	0.111
pHash	0.505	0.316	0.493	0.025
SSIM	0.491	0.305	0.495	0.025
CleanPatrick: Label Errors				
SelfClean	0.572	0.265	0.804	0.018
CLearning	0.477	0.216	0.494	0.002
FINE	0.469	0.205	0.396	0.001
BHN	0.547	0.239	0.545	0.002



(*a*) Kendall's τ (*b*) Spearman's ρ

Figure 5: Correlation coefficient of model rankings produced by both ordinary and soft metrics averaged over 1,000 bootstrapped samples, which represent random annotation draws.

Table 3: Performance comparison of end-to-end fine-tuned models on CheXpert tasks. Rank-flipping is observed (e.g., VGG-16 vs. ResNet-50 on Cardiomegaly AP/s-AP).

Model	AUROC	s-AUROC	AP	s-AP
Cardiomegaly				
EfficientNet-b0	0.752	0.683	0.531	0.480
ResNet-50	0.801	0.711	0.552	0.514
VGG-16	0.782	0.718	0.496	0.493
ViT-base	0.686	0.680	0.407	0.450
Pneumothorax				
EfficientNet-b0	0.206	0.442	0.012	0.042
ResNet-50	0.443	0.544	0.018	0.040
VGG-16	0.510	0.552	0.029	0.051
ViT-base	0.399	0.370	0.016	0.028
Pleural Effusion				
EfficientNet-b0	0.831	0.789	0.556	0.559
ResNet-50	0.928	0.854	0.695	0.669
VGG-16	0.912	0.858	0.601	0.605
ViT-base	0.688	0.669	0.321	0.370

papers. For hyperparameters, we follow the same strategy as *CleanPatrick* and use all methods as implemented in the benchmark (Gröger et al., 2025).

E.1. Off-topic samples

Isolation Forest (IForest): Unsupervised anomaly detection by randomly partitioning features, where points with short average path length are flagged as outliers (Liu et al., 2008).
ECOD: Tail-sensitive, distribution-free outlier scoring based on empirical CDFs computed per feature (Li et al., 2022).
DeepSVD: One-class deep anomaly detection that learns compact representations and flags samples far from the learned support (deep one-class objective) (Ruff et al., 2018).

E.2. Near duplicates

pHash: Perceptual hashing; small Hamming distance between hashes indicates visually similar images (Zauner, 2010).
SSIM: Local comparison of luminance, contrast, and structure, where high mean SSIM signals near-duplicates (Wang et al., 2004).

E.3. Label errors

Confident Learning (CLearning): Estimates class-conditional noise from model predictions to identify likely mislabeled examples (Northcutt et al., 2021).

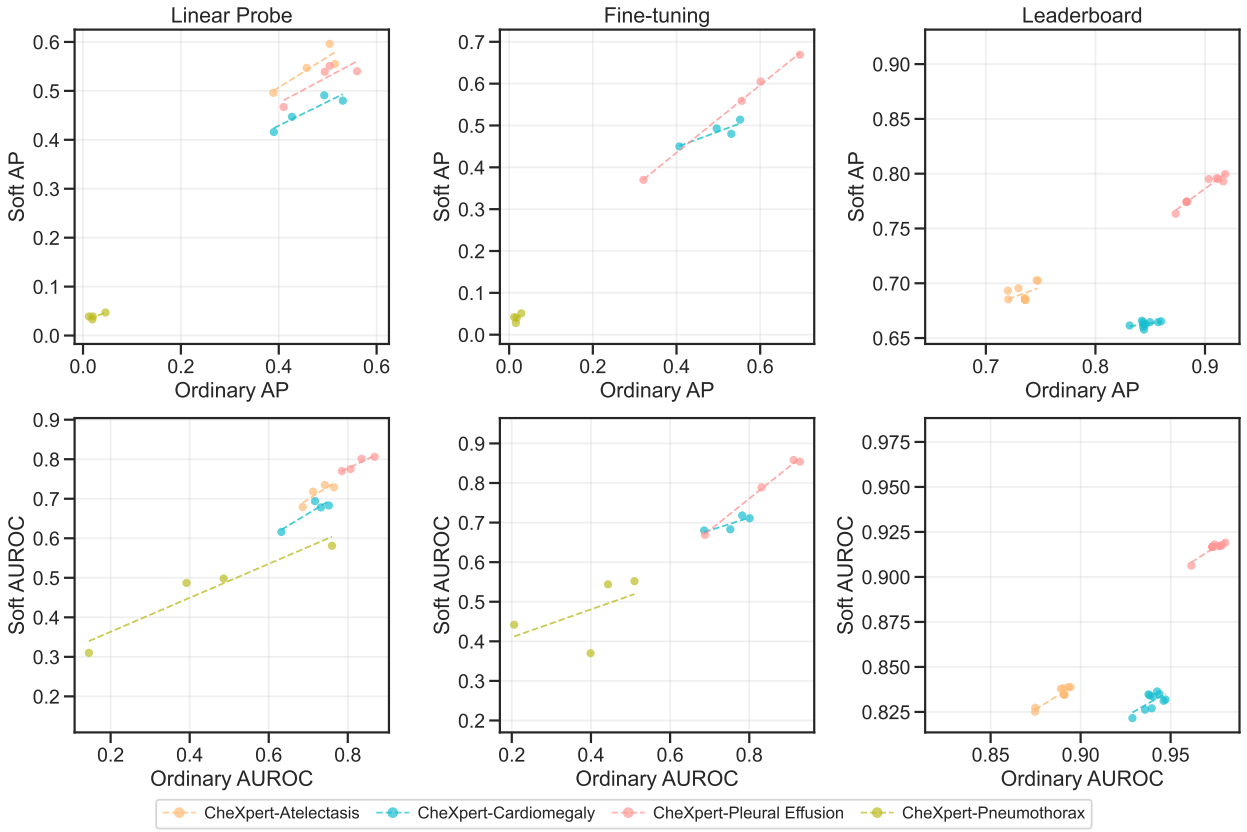


Figure 6: Comparison of ordinary and soft metrics on CheXpert for different tasks. Evaluation is done for linear probing, fine-tuning, and leaderboard models. Dots represent different backbones, and dashed lines indicate Pearson score correlations.

FINE: Influence-based ranking that prioritizes examples whose labels appear inconsistent with the learned decision boundary (Kim et al., 2021).
BHN: Bayesian uncertainty-driven scoring that flags candidates for relabeling based on predictive uncertainty (Yu et al., 2023).

E.4. Multiple issue types

SelfClean: Uses context-aware self-supervised embeddings and neighborhood consistency to rank likely issues (Gröger et al., 2024). The proposed method can be used with a human-in-the-loop or thresholded for automation. Here full automation is used.

Table 4: Re-ranking of CheXternal leaderboard models when switching from ordinary to soft metrics.

	AUROC		**s-AUROC**		**AP**		**s-AP**	
Model	Score	O. Rank	Score	O. Rank	Score	O. Rank	Score	O. Rank
Atelectasis								
jfaboy	0.894	1	0.839	2	0.747	1	0.702	2
ngango3	0.893	2	0.839	1	0.747	2	0.703	1
uestc	0.891	3	0.835	5	0.736	3	0.685	7
drnet	0.891	3	0.835	5	0.736	3	0.685	7
sensexdr	0.891	3	0.835	5	0.736	3	0.685	7
ihil	0.891	3	0.835	5	0.736	3	0.685	7
ngango2	0.890	7	0.838	3	0.730	8	0.696	3
hieupham	0.889	8	0.838	4	0.720	10	0.693	4
desmond	0.875	9	0.827	9	0.720	9	0.685	6
yww211	0.875	10	0.825	10	0.735	7	0.687	5
Cardiomegaly								
sensexdr	0.947	1	0.832	6	0.599	10	0.665	2
ihil	0.946	2	0.831	7	0.857	1	0.665	3
ngango3	0.944	3	0.835	2	0.844	7	0.664	5
hieupham	0.943	4	0.836	1	0.851	2	0.666	1
desmond	0.940	5	0.834	5	0.844	6	0.662	7
drnet	0.940	6	0.827	8	0.844	8	0.660	9
ngango2	0.939	7	0.834	4	0.840	9	0.662	8
yww211	0.938	8	0.835	3	0.850	3	0.665	3
uestc	0.936	9	0.826	9	0.846	4	0.663	6
jfaboy	0.929	10	0.822	10	0.844	5	0.658	10
Consolidation								
jfaboy	0.927	1	0.853	1	0.451	1	0.402	1
uestc	0.921	2	0.848	3	0.408	2	0.397	2
drnet	0.921	2	0.848	3	0.408	2	0.397	2
sensexdr	0.921	2	0.848	3	0.408	2	0.397	2
ihil	0.921	2	0.848	3	0.408	2	0.397	2
yww211	0.918	6	0.851	2	0.354	7	0.392	6
ngango3	0.912	7	0.846	8	0.382	6	0.384	7
desmond	0.905	8	0.847	7	0.330	8	0.383	8
ngango2	0.892	9	0.835	9	0.279	10	0.353	10
hieupham	0.891	10	0.831	10	0.304	9	0.357	9
Edema								
drnet	0.935	1	0.871	1	0.706	1	0.633	1
sensexdr	0.935	1	0.871	1	0.706	1	0.633	1
ihil	0.935	1	0.871	1	0.706	1	0.633	1
uestc	0.935	4	0.869	4	0.694	4	0.628	4
yww211	0.932	5	0.865	5	0.633	8	0.598	7
desmond	0.929	6	0.859	8	0.637	6	0.597	8
ngango2	0.926	7	0.864	6	0.646	5	0.606	5
hieupham	0.924	8	0.862	7	0.637	7	0.603	6
jfaboy	0.922	9	0.856	10	0.566	10	0.566	10
ngango3	0.911	10	0.857	9	0.590	9	0.588	9
Pleural Effusion								
hieupham	0.979	1	0.920	1	0.908	4	0.796	1
ngango3	0.979	2	0.918	3	0.911	3	0.796	2
jfaboy	0.978	3	0.917	4	0.917	1	0.793	5
ngango2	0.977	4	0.917	5	0.350	10	0.795	3
yww211	0.974	5	0.918	2	0.912	2	0.795	3
uestc	0.973	6	0.917	6	0.883	5	0.774	6
drnet	0.973	6	0.917	6	0.883	5	0.774	6
sensexdr	0.973	6	0.917	6	0.883	5	0.774	6
ihil	0.973	6	0.917	6	0.883	5	0.774	6
desmond	0.962	10	0.906	10	0.849	9	0.764	10