Radiology: Artificial Intelligence

SPECIAL REPORT

Hurdles to Artificial Intelligence Deployment: Noise in Schemas and "Gold" Labels

Mohamed Abdalla, PhD • *Benjamin Fine, SM, MD, FRCPC*

From the Institute for Better Health, Trillium Health Partners, Mississauga, Ontario, Canada (M.A., B.F.); and Centre for Information Technology, Department of Computer Science (M.A.), and Department of Medical Imaging (B.F.), University of Toronto, 40 St George St, Room 4283, Toronto, ON, Canada M5S 2E4. Received March 24, 2022; revision requested May 18; revision received November 30; accepted December 19. **Address correspondence to** M.A. (email: *msa@cs.toronto.edu*).

M.A. supported by a Vanier Scholarship from the Government of Canada and a research award from the Vector Institute. The AI Deployment and Evaluation Laboratory at Trillium Health Partners and B.F. are supported by TD Bank, Canada's Supercluster and Trillium Health Partners Foundation.

Conflicts of interest are listed at the end of this article.

See also the commentary by Ursprung and Woitek in this issue.

Radiology: Artificial Intelligence 2023; 5(2):e220056 • https://doi.org/10.1148/ryai.220056 • Content code: AI

Despite frequent reports of imaging artificial intelligence (AI) that parallels human performance, clinicians often question the safety and robustness of AI products in practice. This work explores two underreported sources of noise that negatively affect imaging AI: *(a)* variation in labeling schema definitions and *(b)* noise in the labeling process. First, the overlap between the schemas of two publicly available datasets and a third-party vendor are compared, showing there is low agreement (<50%) between them. The authors also highlight the problem of label inconsistency, where different annotation schemas are selected for the same clinical prediction task; this results in inconsistent use of medical ontologies through intermingling or duplicate observations and diseases. Second, the individual radiologist annotations for the CheXpert test set are used to quantify noise in the labeling process. The analysis demonstrated that label noise varies by class, as agreement was high for pneumothorax and medical devices (percent agreement > 90%). Among low agreement classes (pneumonia, consolidation), the labels assigned as "ground truth" were unreliable, suggesting that the result of majority voting is highly dependent on which group of radiologists is assigned to annotation. Noise in labeling schemas and gold label annotations are pervasive in medical imaging classification and affect downstream clinical deployment. Possible solutions (eg, changes to task design, annotation methods, and model training) and their potential to improve trust in clinical AI are discussed.

Supplemental material is available for this article.

Increasingly, studies report imaging artificial intelligence (AI) algorithms achieving human performance (1–3). However, clinicians who understand the complexity of model development are hesitant to adopt these algorithms in the clinical setting due to questions of reliability, generalizability, model bias, and cognitive biases that may affect safety and robustness in deployment (4), which have all been explored previously (5). In this article, we use the chest radiograph classification task to explore two additional reasons for this hesitancy that have not been well described in the literature but are critical for clinicians to understand to ensure reliable AI model deployment: *(a)* variation in schema definitions and *(b)* noise in the labeling process. Our work describes how these underreported problems, which result from design decisions at the earliest stages of model development but have critical cascading effects on perceived model performance and local deployment (4), can be mitigated. We frame this work from the perspective of a practice looking to deploy an already developed chest radiograph classifier for the task of triage and/or assisted diagnosis. Thus, we focus on the effects of schema and label noise during testing and evaluation and their impact on our understanding of model performance rather than how label noise in training sets affects model training.

Defining Gold Label

Depending on the field of study, the synonymous terms *gold labels*, *gold standard*, *reference standard*, or *ground truth* are used interchangeably. In this article, we purposefully use the term *"gold" label* for two reasons: *(a)* including *label*, which is omitted from other terms, is critical to convey the fact that the labeling process is only an approximation of the "truth" or true standard, and *(b)* the term *gold* is used in quotations to stylistically highlight that label quality might be lower than expected.

Schema Noise

One major diagnostic task of a radiologist interpreting a chest radiograph is to identify one of more than 200 findings that may be present at imaging (6). However, schemas are almost universally limited to about 12 classes (7,8), understandably because of impracticality of annotating hundreds of findings. As a result, binary classification models that are trained on such schemas exclude serious but uncommon findings, such as pneumomediastinum or bone metastases (9), which may result in potentially high-risk, false-negative results.

Schema noise is introduced when different models use different schemas for what users expect to be the same clinical task (ie, chest radiograph abnormality detection). This noise leads to evaluation and deployment challenges, as it is not possible to directly compare the real-world performance of two algorithms that do not have the same set of classes and do not define their classes in the same way.

Abbreviations

ACR = American College of Radiology, AI = artificial intelligence, DSI = Data Science Institute

Summary

By exploring chest radiograph classification as a use case, this report describes the underreported issues of schema and label noise associated with gold label annotations in medical imaging artificial intelligence.

Keywords

Radiology AI, Dataset Creation, Noise in Datasets

Schema Noise Demonstration

We performed a direct comparison of the annotation schemas of two publicly available (7,8) and one proprietary chest radiograph classifier. In addition to quantifying the magnitude of overlap between these schemas, we also leveraged existing radiologic ontologies (6) to highlight (the lack of) conceptual agreement within individual schemas. We define schema overlap as the fraction of shared labels between two schemas out of the total number of unique labels.

Table S1 demonstrates that the overlap between schemas varies greatly: CheXpert and ChestX-ray14 demonstrate 37% schema overlap (seven shared labels out of 19 unique labels between the two schemas); CheXpert and the proprietary classifier show 28% schema overlap (five of 18); and ChestX-ray14 and the proprietary classifier show 35% schema overlap (six of 17).

Further exploring the schema of these chest radiograph algorithms, we find multiple sources of noise related to inconsistent class definition both across models and within each model. Table 1 presents a noncomprehensive list of sources of schema noise, including class overlap (infiltrate vs consolidation), hidden hierarchy (abscess as a leaf node of cavity), and intermingling observations and disorders (consolidation vs pneumonia), with implications on model development and radiologist agreement. For example, Figure 1 displays part of the Radiology Gamuts Ontology, where labels (outlined in black) belong to both "disorders" and "observations caused by disorders," illustrating inconsistency in schema design.

Label Noise

In addition to (and sometimes as the consequence of) schema noise, disagreement between radiologists (ie, high interobserver variability) introduces label noise. Radiologists are known to disagree when interpreting chest radiographs (10,11). Yet, current practice is to train models on binary labels created by a panel of radiologists which, often after a voting process, are collapsed to a single (often binary) output. We show how failure to report this uncertainty (hidden by the use of binary or categorical classes) negatively affects users who aim to evaluate model performance. Specifically, using the CheXpert annotation dataset, we illustrate variations in the gold label depending on which radiologists are included in the panel, suggesting that binarization via majority vote does not sufficiently increase reliability.

Label Noise Demonstration

Label-Level Agreement

To demonstrate the level of noise visible in trusted gold labels, we simulated the creation of different gold label sets using CheXpert test annotations provided by the authors of the CheXpert article, which includes further details for this dataset (eg, information about annotators) (12). No human research was performed; therefore, this study was exempt from institutional review board review. Originally, the creators of CheXpert had eight annotators label 500 images for the 14 different classes and defined the gold label as the majority vote of five annotators chosen randomly from the eight. The remaining three radiologists were used to establish a performance baseline for the AI models. To demonstrate how reliant the gold label set—and therefore the performance evaluations—would be from the five annotators chosen, we created all possible annotation panels of five radiologists by sampling (without replacement) from the set of annotations from the eight radiologists (Fig 2), resulting in $\binom{8}{5}=56$ possible gold label sets for each label. Each of these sets of five is a possible gold label set in the real world (that could have resulted from the random selection). We then calculated the agreement of all pairings of possible gold label sets using select statistical measures to measure agreement between observations.

Median percent agreement.—This common measure of agreement is intuitive as it simply measures the number of concordant observations (between two observers) over the total number of observations. Because we compare all possible pairwise pairings of potential gold label sets, we report the aggregate of all these comparisons using the median value, which is more robust to outliers.

Median Cohen κ.—This metric is commonly used in the medical literature for assessing the agreement between two labels (or raters). It differs from percent agreement in that Cohen κ also corrects for the probability that the agreement could have occurred by chance. This is important in radiograph classification because most classes are not present in most examinations, increasing the likelihood of chance agreement when findings are not present. The mathematical definition of Cohen κ is as follows:

$$\kappa = \frac{p_0 - p_e}{1 - p_e},$$

where p_0 is the observed agreement and p_e is the probability of chance agreement. Like above, we opt to report the median κ, as the median is more robust to outliers.

Fleiss κ.—Unlike the above two measures, Fleiss κ can compare more than two raters. Like Cohen κ, it accounts for chance agreement between all raters (13).

Table 1: Possible Sources of Noise in Schema Design, Both within a Schema and across Different Schemas

Source of Noise	Example of Noise	Implications
Intraschema		
Class overlap	Infiltrate and consolidation are overlapping concepts.	A group of radiologist labelers are unlikely to label consistently.
Hidden hierarchy	Cavity is a parent node of abscess.	A group of radiologist labelers are unlikely to label consistently.
Inclusion of both findings (observations) and diagnosis (disorders)	Consolidation is a radiographic finding that, in the context of an otherwise healthy patient with cough and fever, may represent the disorder pneumonia (Fig 1).	A group of radiologist labelers are unlikely to label consistently.
Interschema		
Inconsistent inclusion and exclusion of findings	Only one of the algorithms explored is trained to detect fractures.	Creates cognitive load and/or uncertainty in the minds of busy clinical users: Which abnormalities might still be present if the algorithm is negative?

We observed different levels of agreement across labels: Median agreement was high (0.95) for support devices and low (0.65) for labels such as consolidation and pneumonia (Fig 3). The statistical methods and complete results are presented in Table S2.

For a more intuitive understanding of label noise, we also present the median percent agreement for positive (ie, abnormal) cases between all possible pairings of simulated gold label sets for each diagnosis. As most labels are negative (ie, normal), percent agreement on the entire dataset is not informative. As illustrated in Figure 3, for the three labels with the lowest agreement (consolidation, fracture, and pneumonia), median agreement between different panels of radiologists for positive examples is equivalent to a series of coin flips. Fleiss κ values were low to medium and showed general discordance between multiple annotators, demonstrating that a single annotator is not responsible for the noise in the labels.

Patient-Level Agreement (Distribution across Patients)

To better understand the source of label noise, we explored how disagreement in gold labels is distributed across patients by counting the number of labels per patient that have at least two (or three or four) radiologists disagreeing. This will help determine if disagreement is randomly distributed among all patients or if most of the label noise (disagreement) originates from a small subset of patients, thus directing efforts to explore and address the causes of noise.

We analyzed patient-level disagreements in the CheXpert annotation dataset (eight radiologists, 14 classes, 500 patients) and present the results in Figure 4. Figure 4A shows broad annotator disagreement; at least two radiologists disagreed on one or more labels (out of 14) in more than 90% (460 of 500) of patients. That is, there was complete agreement among all radiologists on all classes in fewer than 10% of cases (40 of 500). For most patients, 25% (two of eight) annotator disagreement was present for three or more labels. Figure 4C shows the extreme case where radiologists are split 50:50 (four of eight) on one label. In 40% of patients (201 of 500), half of annotating radiologists disagreed on at least one label. In other words, the assigned gold label for at least one label in 40% of patients could equivalently be decided by a coin flip. This analysis demonstrates the magnitude of noise that is hidden in the gold labels by use of collapsed binary labels per image. We highlight the importance of describing this noise and review methods of countering label noise in the discussion that follows.

Importance of Identifying Schema and Label Noise

Past work has explored noise in the training sets of computer vision models resulting from lack of annotator agreement (in AI radiology [10,14] and other applications [15,16]), as well as the imperfect performance of automated natural language processing methods (17). Label noise in training data has been demonstrated to have negative effects on machine learning performance (17–19), though substantial literature demonstrates how this effect can be ameliorated and how noisy labels can be used to improve machine learning performance by increasing variance in the data and acting as a regularizer (17,20–22). There are also post hoc approaches to manage label noise, including label cleaning (17) and smoothing (3), changing network architectures (17,23), and training strategies and/or pipelines (20,23,24).

However, in the clinically important setting of external evaluation (eg, where radiologists are exploring the possible deployment of third-party models), noise (which can translate to errors) in the gold label set can limit user understanding of true model performance. This hurdle to AI deployment is rarely addressed and is the main concern of this work.

Our article can be considered as an exploration of "data cascades" (4) in the field of radiology. Data cascades are choices related to problem definitions, data collection and labeling, model

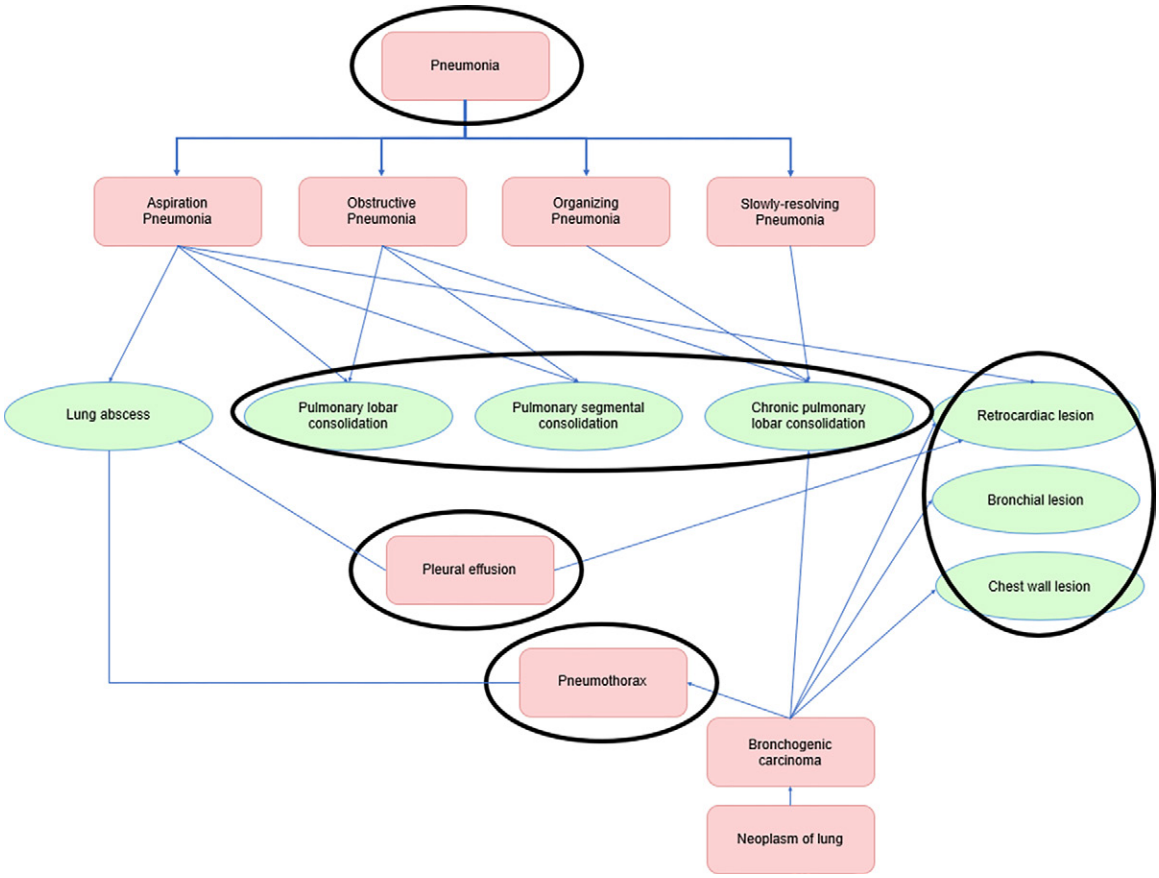

Figure 1: Visualization of part of the Radiology Gamuts Ontology (6). Red rounded rectangles represent disorders, while green ovals represent observations caused by disorders. Labels (either red or green) that are in the CheXpert schema are outlined in thick black. This figure illustrates how such a schema will result in researchers and developers designing a model that lacks clarity on prediction classes required for clinical deployment use case and intermingles the prediction of both observations and disorders, which makes interpretation of model predictions, and thus translation into clinical practice, challenging.

selection, and other machine learning development steps, which cause negative downstream effects that compound on each other. Many of the technical solutions proposed to reduce or deal with noisy training labels are designed for model selection and training; these statistical approaches cannot be easily applied to evaluation of models (internal or external testing). Other work for improving radiologist agreement (eg, advanced annotation frameworks [25]) is difficult (if not impossible) to apply post hoc to available data. Unlike past work, we focus on highlighting critical problems and corresponding solutions related to data cascades that occur earlier (ie, problem definition, data labeling) and later (ie, local validation) in the machine learning deployment pipeline.

Noise Mitigation Strategies

Reducing Noise

Tables 2 and 3 present a noncomprehensive list of various schema and label noise mitigation strategies. Some of the proposed strategies are the responsibility of individual researchers or developers (eg, reporting certain metrics or asking certain questions before data collection), while others require community-led efforts (eg, developing consensus schemas).

Anchoring Labels in Ontologies

To address the inconsistency in schema designs, the AI community can build atop of existing large-scale projects defining radiologic ontologies and knowledge graphs of medical concepts, such as the Radiology Gamuts Ontology (6). These knowledge graphs provide a comprehensive list of disorders and observations, which are linked by ontological relationships (6). The structure of these knowledge graphs enables schema designers to maintain consistency in their schema designs (Fig 1). As demonstrated in a recent work (26), we believe leveraging ontology-based schemas can help bridge the gap to clinical deployment by enabling researchers and developers to design better label schemas.

At the same time, reviewers and end users should insist on schemas that reflect ontological knowledge and fit real-world use cases—applying post hoc fixes or transformations on datasets to enable end-use application is much more expensive and time-consuming (a concept termed *technical debt* in machine learning [27]).

In addition to individualized efforts via reviewers and users, informatics societies could increase efforts to define, and push the adoption of, standardized annotation schemata. The American College of Radiology (ACR) Data Science Institute

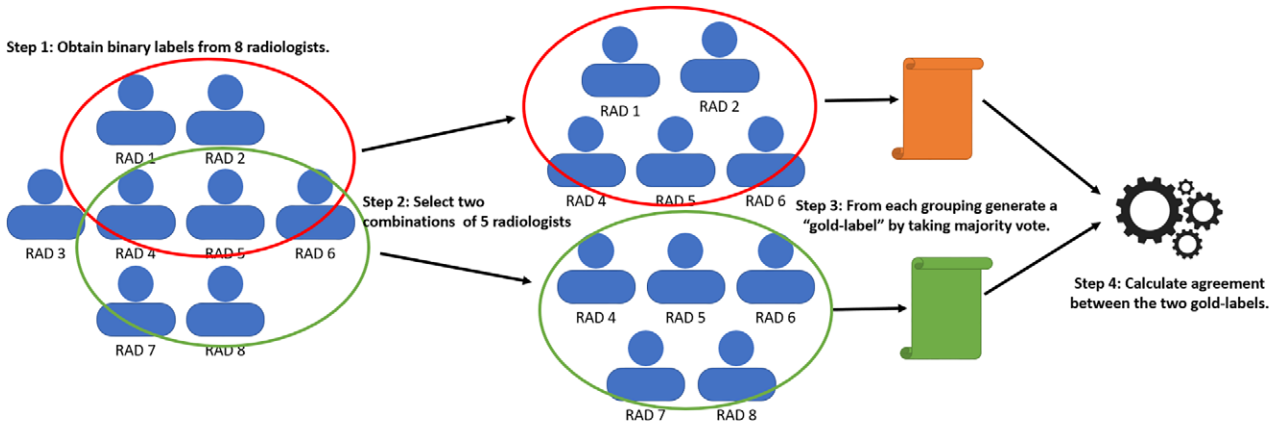

Figure 2: Representation of the simulation method measuring variation in the annotation process of a chest radiograph annotation task. We simulated labels created by majority vote when labels are provided by a randomly selected group of five radiologists out of the original eight labeling radiologists. We then measured agreement between all gold labels resulting from all possible gold label sets ($n = 56$) in a pairwise fashion. RAD = radiologist.

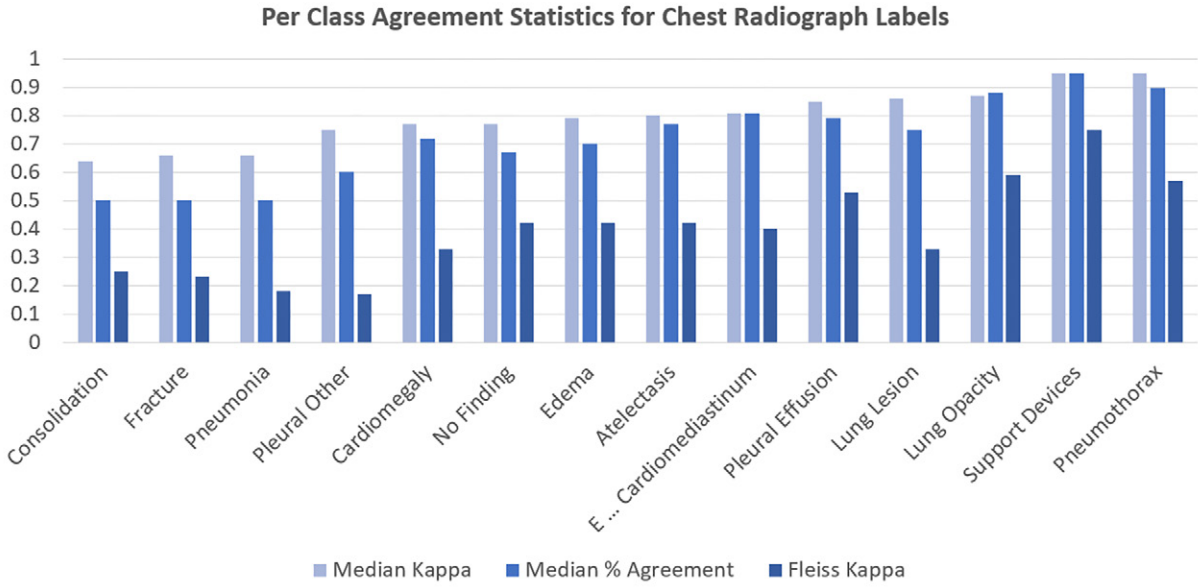

Figure 3: The median agreement (measured using Cohen κ and percent agreement of positive cases) between all pairwise comparisons of all possible gold label sets, sorted in ascending order. We also calculated Fleiss κ—a statistical measure for assessing reliability of agreement between multiple raters at once. A table of values can be found in Tables S1 and S2. E = enlarged.

(DSI) has taken steps in this direction by introducing label schemas for specific use cases (discussed below).

Standardizing Use Cases—Define AI Use Case Directory via ACR DSI

Disagreements in schema design are often the result of differences in local practice and radiologist or scientist subjectivity. Efforts to standardize AI use cases such as that of the ACR DSI (28) can help reduce noise and improve generalizability in multiple ways. First, adoption of these use cases can result in the standardization of schema, preventing the current issues with similar yet noninteroperable schemas. This would facilitate direct comparison of AI models for the same clinical task, aiding the efforts of researchers to improve on state-of-the-art models, of imaging professionals to evaluate and deploy AI, and of third-party vendors trying to instill confidence in their products.

Describing Label Noise

To reduce label noise, authors' and developers' description of their data collection and annotation generation methods should include a measurement of noise in the labels (#18 of CLAIM checklist [29]). Such evaluation will not only help developers understand the performance of models trained on these datasets, but also help clinicians gain trust in the performance of said models. Data scientists should also recognize when the majority-

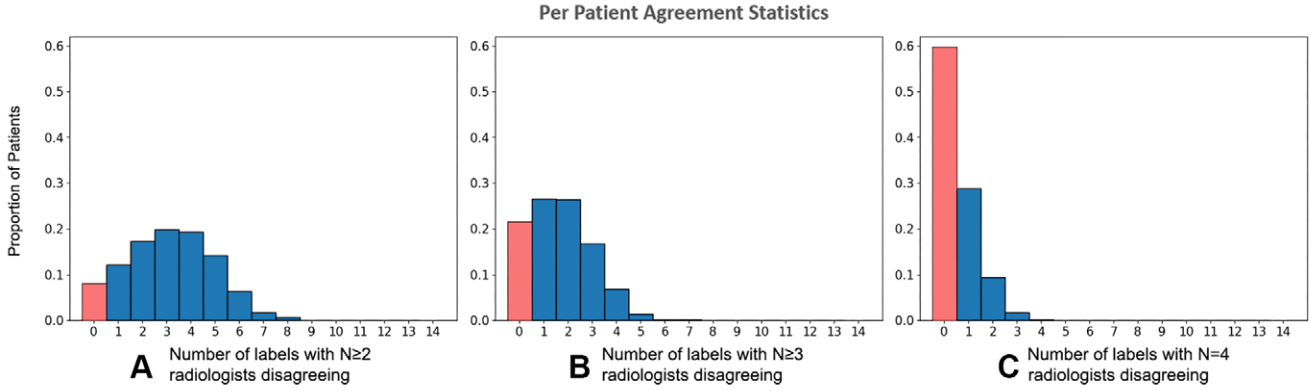

Figure 4: A histogram counting the percent of patients in the CheXpert gold label–annotation dataset where at least *n* = 2,3,4 radiologists disagree from the assigned gold label **(A–C)**, binned by number of labels (of 14) demonstrating the disagreement. The first (0) bar, in red, represents the number of patients who have no labels with at least *n* = 2,3,4 radiologist annotator disagreement. For example, in **A**, only about 10% of patients (40 of 500) had no classes where two radiologists disagreed—indicating that only 10% of cases showed complete agreement across classes by all radiologists. The remaining cases all demonstrated noise (disagreement) in assigned labels, which is hidden in "ground truth labels."

Table 2: Various Strategies to Mitigate Schema Noise

Mitigation Strategy	Examples and Citations
Design schemas that explicitly fit intended use	The American College of Radiology Data Science Institute has proposed the introduction of standardization of use cases as a starting point (28). Without such standardization, similar yet noninteroperable schema will proliferate, making comparison and analyses of proposed artificial intelligence algorithms difficult and disorderly. New schemas, if more appropriate to develop and/or use for the intended use case, should be explicitly justified.
Provide a datasheet for the schema and the dataset	Gebru et al's datasheets for datasets proposes the introduction of a standardized process for documenting dataset (30). See example datasheet for CheXpert dataset datasheet (12). We propose including schema design justification and label noise description.
Develop consensus schemas	Currently, dataset schemas are created to maximize the usage of available data from any specific institution. While this is a reasonable approach, over time this approach will lead the field into a variety of competing and noninteroperable standards. To address this, there should be standardization of schemas. For real-world use cases, researchers and practicing radiologists should collaborate to develop a fully specified schema that fulfills the clinical requirement for the specific use case.
Incorporate existing knowledge by using hierarchical schemas or graph schemas for labeling related observations and diagnoses	A body of literature describes medical ontologies that should be used to label schema design. For example, *www.gamuts.net* is a knowledge graph linking findings and disorders across the entire spectrum of radiology imaging (6), which can be used to design labeling schemata.

vote "hard" labeling methods produce a false sense of certainty and noise should instead be represented using "soft" labels (18).

Label noise has already been described as a component of datasheets for datasets that can facilitate communication of annotation between researchers, developers, and end users. In Section 3.2 titled "Composition," Gebru et al (30) highlight questions regarding the composition of a dataset that should be answered by the data holders. Among these questions are ones regarding sources of noise or errors within the dataset: "Are there any errors, sources of noise, or redundancies in the dataset? If so, please provide a description."

Garbin et al (12) applied the datasheets for datasets concept to the CheXpert dataset; however, label noise is not explicitly quantified or addressed. To address this, we suggest that researchers consider these two radiology-specific questions: *(a)* What is the use case or justification behind the schema design? and *(b)* How accurate and reliable are your gold labels? Have you quantified both measures? Addressing these questions can reduce schema and label noise and enable more informed decisions by downstream users of such datasets for model development.

One place to apply this label noise description for great potential impact would be open data science challenges (eg, the yearly open tasks hosted by the Radiological Society of North America [31]). It would push the field forward if organizers of the yearly event reported both the justification of the annotation schema and the label noise measures highlighted in this article. Publicly describing the schema justification and reporting label noise would help expose many researchers to the issues discussed

Table 3: Various Strategies to Reduce or Adapt to Label Noise

Strategy	Examples and Citations
Report label noise and internal consistency	Currently, most gold labels are presented as hard labels without any measure of uncertainty or noise. Future work should report statistics about annotation noise and consistency to help developers and clinicians understand when such labels should be treated as soft labels.
Labeling method should be selected to match the intended use case	Currently, the most commonly used labeling method is majority vote. However, depending on the intended use case, this may not be the most appropriate: *(a)* For example, in a screening use case, it may be worthwhile to explore using another labeling method that maximizes sensitivity (eg, positive label if any of the radiologists labeled it as positive). *(b)* Where agreement is desired for the specific use case, we recommend adjudicative labeling methods to improve agreement (25). *(c)* Where the use case requires accurate ordering or ranking of cases in terms of severity, we recommend using comparative annotations. Comparative annotations can be used to provide an ordinal ranking of images by severity (32). Comparative annotations have been used in the labeling of other machine learning datasets (32,33) and provides solutions to issues of traditional labeling methods (32,34).
Incorporate more definitive follow-up tests to validate labels	Chest radiograph labels produced by radiologists are often noisy and not suitable for the creation of a ground truth label. We recommend attempting to incorporate other information to generate gold labels. For example, depending on the intended use case, researchers can use CT images acquired contemporaneously at the time of a radiograph as a source of radiograph labels. Alternatively, incorporating laboratory or pathologic results could be used to help determine true labels in some clinical scenarios.
Use noise-tolerant training methods	A review of methods of adapting architecture, regularization, loss design, and sample selection for noise is presented in Song et al (18).

in this article. Moreover, public reporting would help normalize the act of measuring and reporting such data. If agreement was reported per image, researchers would be able to develop methods that incorporate this noise during training. This act alone would also improve understanding of acceptable ranges for noise (eg, what is a reasonable amount of label agreement for different tasks?). This reporting can also serve to teach others how noise can be reported for different types of data (eg, percent agreement or κ for categorical labels and different approaches such as intersection over union for bounding box approaches).

Study Limitations

There were limitations to our work that provide opportunity for future research efforts. First, our study was limited to the test set of CheXpert data (due to availability); findings may not generalize to other imaging modalities. We hope that in the future, those releasing datasets are more open to sharing their annotations with researchers or perform and publish this type of analysis on their test set. Second, this work does not examine the images underlying the annotations to determine the predominant sources of disagreement. For such an analysis, we direct readers to the work of Duggan et al (25). Third, our formulation of label uncertainty does not actively consider the uncertainty of the individual annotators; the dataset did not contain such labels. Knowing the uncertainty of individual annotators can provide meaningful insight to label adjudication: Three radiologists who are 90% confident there is a fracture who disagree with two radiologists who are only 50% confident there is no fracture would be treated differently if all radiologists were only 50% confident of their decision.

Another limitation, stemming from the underlying dataset, was the fact that the task at hand was focusing solely on image interpretation. In the real world, clinical history and prior images would be present, possibly leading to lower disagreements. However, as most chest radiograph AI models are trained from images alone, we felt it was appropriate to use these data for analysis. Finally, we chose to highlight the agreement between multiple pairwise comparisons rather than a metric that compares all annotators at once (Fleiss κ) as it would allow for a more intuitive comparison between pairs of gold label sets. However, we do also present group metrics for interested researchers. Regardless of the measure used, there is a clear need to understand how uncertainty interacts with trained models and its impact on clinical outcomes.

Conclusion

In summary, underreported noise in schema design and gold labels are widespread and contribute to mistrust and challenges in development, evaluation, and clinical deployment of chest radiograph AI models.

Our work *(a)* describes the concept of schema noise—the lack of conceptual clarity regarding the task at hand, manifesting itself as class overlap (infiltrate vs consolidation), hidden hierarchy (abscess as a leaf node of cavity), and intermingling observations with disorders (consolidation vs pneumonia) and *(b)* characterizes the second-order variation and quantifies the magnitude of label noise.

Our findings emphasize the considerable amount of noise in gold labels used to evaluate models, which is critical for end users to understand in the context of evaluating the safety and robustness of externally developed algorithms for local deployment. We describe schema and label noise to *(a)* help clinical users answer the question, "How much can I trust a positive or negative result from this chest radiograph classifier?" and *(b)* guide developers to recognize and mitigate this noise. Both researchers and organizations have a role to play in mitigating noise through transparent justification of decision-making and reporting of noise to prevent unintended downstream degradation of AI performance and trust.

Acknowledgments: Thank you to Pranav Rajpurkar, PhD, and Matthew Lungren, MD, MPH, for sharing the CheXpert annotation dataset.

Author contributions: Guarantors of integrity of entire study, **M.A., B.F.**; study concepts/study design or data acquisition or data analysis/interpretation, **M.A., B.F.**; manuscript drafting or manuscript revision for important intellectual content, **M.A., B.F.**; approval of final version of submitted manuscript, **M.A., B.F.**; agrees to ensure any questions related to the work are appropriately resolved, **M.A., B.F.**; literature research, **M.A., B.F.**; clinical studies, **B.F.**; experimental studies, **B.F.**; statistical analysis, **M.A., B.F.**; and manuscript editing, **M.A., B.F.**

Data availability: The data were obtained without right of redistribution. Researchers can request access to the data from can be requested from the authors of CheXpert (8).

Disclosures of conflicts of interest: M.A. No relevant relationships. **B.F.** Support from TD Ready Commitment, Digital Supercluster Canada, Trillium Health Partners Foundation; grants/contracts from University of Toronto; consultancy fees from Ascertain; participation on a Data Safety Monitoring Board or Advisory Board for Canon Medical; role on SIIM Machine Learning Committee and HaloHealth; stock/stock options in Elevens AI, Eva Medical, HeartVista, PocketHealth, Iterative Health.

References

1. Rajpurkar P, Irvin J, Ball RL, et al. Deep learning for chest radiograph diagnosis: A retrospective comparison of the CheXNeXt algorithm to practicing radiologists. PLoS Med 2018;15(11):e1002686.
2. Rajpurkar P, Irvin J, Zhu K, et al. CheXNet: Radiologist-Level Pneumonia Detection on Chest X-Rays with Deep Learning. arXiv 1711.05225 [preprint] https://arxiv.org/abs/1711.05225. Posted November 14, 2017. Accessed September 2021.
3. Pham HH, Le TT, Tran DQ, Ngo DT, Nguyen HQ. Interpreting chest X-rays via CNNs that exploit hierarchical disease dependencies and uncertainty labels. Neurocomputing 2021;437:186–194.
4. Sambasivan N, Kapania S, Highfill H, Akrong D, Paritosh P, Aroyo LM. "Everyone wants to do the model work, not the data work": Data Cascades in High-Stakes AI. In: Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems 2021 May 6; 1–15.
5. Asan O, Bayrak AE, Choudhury A. Artificial intelligence and human trust in healthcare: focus on clinicians. J Med Internet Res 2020;22(6):e15154.
6. Budovec JJ, Lam CA, Kahn CE Jr. Informatics in radiology: radiology gamuts ontology: differential diagnosis for the Semantic Web. RadioGraphics 2014;34(1):254–264.
7. Wang X, Peng Y, Lu L, Lu Z, Bagheri M, Summers RM. ChestX-ray8: Hospital-scale chest X-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition 2017; 3461–3471.
8. Irvin J, Rajpurkar P, Ko M, et al. CheXpert: A Large Chest Radiograph Dataset with Uncertainty Labels and Expert Comparison. Proc AAAI Conf Artif Intell 2019;33(1):590–597.
9. Oakden-Rayner L, Dunnmon J, Carneiro G, Ré C. Hidden stratification causes clinically meaningful failures in machine learning for medical imaging. In: Proceedings of the ACM Conference on Health, Inference, and Learning 2020 Apr 2; 151–159.
10. Sakurada S, Hang NT, Ishizuka N, et al. Inter-rater agreement in the assessment of abnormal chest X-ray findings for tuberculosis between two Asian countries. BMC Infect Dis 2012;12(1):31.
11. Neuman MI, Lee EY, Bixby S, et al. Variability in the interpretation of chest radiographs for the diagnosis of pneumonia in children. J Hosp Med 2012;7(4):294–298.
12. Garbin C, Rajpurkar P, Irvin J, Lungren MP, Marques O. Structured dataset documentation: a datasheet for CheXpert. arXiv:2105.03020 [preprint] https://arxiv.org/abs/2105.03020. Posted May 7, 2021. Accessed September 2021.
13. Fleiss JL. Measuring nominal scale agreement among many raters. Psychol Bull 1971;76(5):378–38.
14. Balabanova Y, Coker R, Fedorin I, et al. Variability in interpretation of chest radiographs among Russian clinicians and implications for screening programmes: observational study. BMJ 2005;331(7513):379–382.
15. Northcutt CG, Athalye A, Mueller J. Pervasive Label Errors in Test Sets Destabilize Machine Learning Benchmarks. arXiv:2103.14749 [preprint] https://arxiv.org/abs/2103.14749. Posted March 26, 2021. Accessed May 2022.
16. Yu X, Han B, Yao J, Niu G, Tsang I, Sugiyama M. How does Disagreement Help Generalization against Label Corruption? In: International Conference on Machine Learning 2019 May 24; 7164–7173. PMLR. https://proceedings.mlr.press/v97/yu19b.html.
17. Karimi D, Dou H, Warfield SK, Gholipour A. Deep learning with noisy labels: Exploring techniques and remedies in medical image analysis. Med Image Anal 2020;65:101759.
18. Song H, Kim M, Park D, Shin Y, Lee JG. Learning From Noisy Labels With Deep Neural Networks: A Survey. IEEE Trans Neural Netw Learn Syst 2022. 10.1109/TNNLS.2022.3152527. Published online March 7, 2022.
19. Gatt ME, Spectre G, Paltiel O, Hiller N, Stalnikowicz R. Chest radiographs in the emergency department: is the radiologist really necessary? Postgrad Med J 2003;79(930):214–217.
20. Wu N, Phang J, Park J, et al. Deep neural networks improve radiologists' performance in breast cancer screening. IEEE Trans Med Imaging 2020;39(4):1184–1194.
21. Krause J, Sapp B, Howard A, et al. The unreasonable effectiveness of noisy data for fine-grained recognition. In: European Conference on Computer Vision 2016 Oct 8; 301–320. Springer, Cham. https://doi.org/10.1007/978-3-319-46487-9_19.
22. Sun C, Shrivastava A, Singh S, Gupta A. Revisiting Unreasonable Effectiveness of Data in Deep Learning Era. arXiv preprint arXiv:1707.02968. https://arxiv.org/abs/1707.02968. Posted July 10, 2017.
23. Ju L, Wang X, Wang L, et al. Improving Medical Images Classification With Label Noise Using Dual-Uncertainty Estimation. IEEE Trans Med Imaging 2022;41(6):1533–1546.
24. Gündel S, Setio AAA, Ghesu FC, et al. Robust classification from noisy labels: Integrating additional knowledge for chest radiography abnormality assessment. Med Image Anal 2021;72:10208.
25. Duggan GE, Reicher JJ, Liu Y, Tse D, Shetty S. Improving reference standards for validation of AI-based radiography. Br J Radiol 2021;94(1123):20210435.
26. Jain S, Agrawal A, Saporta A, et al. RadGraph: Extracting Clinical Entities and Relations from Radiology Reports. arXiv:2106.14463 [preprint] https://arxiv.org/abs/2106.14463. Posted June 28, 2021. Accessed October 2022.
27. Sculley D, Holt G, Golovin D, et al. Hidden Technical Debt in Machine Learning Systems. Adv Neural Inf Process Syst 2015; 28. https://papers.nips.cc/paper/2015/hash/86df7dcfd896fcaf2674f757a2463eba-Abstract.html.
28. Allen B. How structured use cases can drive the adoption of artificial intelligence tools in clinical practice. J Am Coll Radiol 2018;15(12):1758–1760.
29. Mongan J, Moy L, Kahn CE Jr. Checklist for artificial intelligence in medical imaging (CLAIM): a guide for authors and reviewers. Radiol Artif Intell 2020;2(2):e200029.
30. Gebru T, Morgenstern J, Vecchione B, et al. Datasheets for datasets. Commun ACM 2021;64(12):86–92.
31. RSNA. AI Challenges. https://www.rsna.org/education/ai-resources-and-training/ai-image-challenge. Published 2017. Accessed August 22, 2022.
32. Kiritchenko S, Mohammad S. Best-Worst Scaling More Reliable than Rating Scales: A Case Study on Sentiment Intensity Annotation. In: Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Vol 2: Short Papers) 2017 Jul; 465–470. https://doi.org/10.18653/v1/P17-2074.
33. Asaadi S, Mohammad S, Kiritchenko S. Big BiRD: A Large, Fine-Grained, Bigram Relatedness Dataset for Examining Semantic Composition. In: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Vol 1 (Long and Short Papers) 2019 Jun; 505–51.
34. Baumgartner H, Steenkamp JB. Response styles in marketing research: A cross-national investigation. J Mark Res 2001;38(2):143–156.