Radiology: Artificial Intelligence

ORIGINAL RESEARCH

The Subgroup Imperative: Chest Radiograph Classifier Generalization Gaps in Patient, Setting, and Pathology Subgroups

Monish Ahluwalia, MD, MSc • Mohamed Abdalla, PhD • James Sanayei, MD • Laleh Seyyed-Kalantari, PhD • Mohannad Hussain, HBSc • Amna Ali, HBSc • Benjamin Fine, SM, MD

From the Kingston Health Sciences Centre, Queen's University, Kingston, Ontario, Canada (M. Ahluwalia); Faculty of Medicine (M. Ahluwalia, J.S.), Institute of Health Policy, Management and Evaluation (M. Ahluwalia), Department of Computer Science (M. Abdalla, L.S.K.), and Department of Medical Imaging (B.F.), University of Toronto, Toronto, Ontario, Canada; Vector Institute for Artificial Intelligence, Toronto, Canada (M. Abdalla, B.F.); Institute for Better Health (M. Abdalla, A.A., B.F.) and Department of Diagnostic Imaging (A.A., B.F.), Trillium Health Partners, 100 Queensway West, Clinical Administrative Building, 6th Floor, Mississauga, ON, Canada L5B 1B8; Department of Medicine, Royal University Hospital, Saskatoon, Saskatchewan, Canada (J.S.); Department of Electrical Engineering and Computer Science, York University, Toronto, Ontario, Canada (L.S.K.); and Techie Maestro, Waterloo, Ontario, Canada (M.H.). Received November 30, 2022; revision requested March 6, 2023; revision received June 6; accepted June 22. **Address correspondence to** M. Ahluwalia (email: *14ma84@queensu.ca*).

Supported by Digital Supercluster Canada.

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

See also the commentary by Huisman and Hannink in this issue.

Radiology: Artificial Intelligence 2023; 5(5):e220270 • https://doi.org/10.1148/ryai.220270 • Content codes: AI CH

Purpose: To externally test four chest radiograph classifiers on a large, diverse, real-world dataset with robust subgroup analysis.

Materials and Methods: In this retrospective study, adult posteroanterior chest radiographs (January 2016–December 2020) and associated radiology reports from Trillium Health Partners in Ontario, Canada, were extracted and de-identified. An open-source natural language processing tool was locally validated and used to generate ground truth labels for the 197 540-image dataset based on the associated radiology report. Four classifiers generated predictions on each chest radiograph. Performance was evaluated using accuracy, positive predictive value, negative predictive value, sensitivity, specificity, F1 score, and Matthews correlation coefficient for the overall dataset and for patient, setting, and pathology subgroups.

Results: Classifiers demonstrated 68%–77% accuracy, 64%–75% sensitivity, and 82%–94% specificity on the external testing dataset. Algorithms showed decreased sensitivity for solitary findings (43%–65%), patients younger than 40 years (27%–39%), and patients in the emergency department (38%–60%) and decreased specificity on normal chest radiographs with support devices (59%–85%). Differences in sex and ancestry represented movements along an algorithm's receiver operating characteristic curve.

Conclusion: Performance of deep learning chest radiograph classifiers was subject to patient, setting, and pathology factors, demonstrating that subgroup analysis is necessary to inform implementation and monitor ongoing performance to ensure optimal quality, safety, and equity.

Supplemental material is available for this article.

Increasing demand for imaging combined with a shortage of radiologists in many regions of the world has led to growing interest in implementing deep learning classifiers into radiology workflows (1–4). Classifiers implemented in these settings must perform safely and equitably.

Commercial and open-source radiology classifiers are increasingly available and have been developed using datasets that are reported to be unbiased with well-established ground truth (5,6). Despite this, decreased performance when implementing classifiers in new clinical settings and previously unseen data, also known as a generalization gap, is well established (7–11). A recent systematic review showed that most external testing studies in radiology demonstrate modestly decreased algorithm performance (12). In addition, most external testing studies are conducted on small datasets, which may not provide sufficient information on the generalizability of these models, especially for underrepresented subgroups or less common abnormalities (13). Subgroup analysis on protected attributes has become a critical task because state-of-the-art chest radiograph classifiers have demonstrated bias against certain patient subgroups (14,15). Although overall performance is important, health systems have a mandate to ensure equity and fairness in health care delivery (16).

Thus, we present a large-scale (n = 197 540) external testing of four state-of-the-art deep learning chest radiograph classifiers (14) (three trained on the three largest publicly available chest radiograph datasets—the CheXpert Image [1], MIMIC-CXR [17], and Chest X-ray-14 [18] datasets—and one proprietary third-party classifier) in a binary classification task. We pursued robust subgroup analysis on patient, setting, and pathology subgroups to provide the best understanding of performance and respect the mandates of clinicians, health system engineers, and policymakers alike.

Abbreviations

DICOM = Digital Imaging and Communications in Medicine, EHR = electronic health record, ICU = intensive care unit, MCC = Matthews correlation coefficient, NLP = natural language processing, NPV = negative predictive value, PPV = positive predictive value

Summary

Chest radiograph classifier performance varied in patient, setting, and pathology subgroups, demonstrating that subgroup analysis is critical during external testing to identify gaps that affect safe and equitable deployment.

Key Points

- Four state-of-the-art chest radiograph classifiers showed lower sensitivity on younger patients (an absolute 33% decrease), emergency department patients (an absolute 12% decrease), and patients with a solitary abnormality (an absolute 27% decrease).
- Receiver operating characteristic analysis showed that subgroup performance can represent different locations on an operating curve, differences that may be obscured by classifiers with single performance metrics.
- Open-source natural language processing tools performed adequately (accuracy of 94%, precision and sensitivity > 90%) as a scalable means to extract ground truth from chest radiograph reports.

Keywords

Conventional Radiography, Thorax, Ethics, Supervised Learning, Convolutional Neural Network (CNN), Machine Learning Algorithms

Materials and Methods

Study Design

This study was approved by the Research Ethics Board at Trillium Health Partners, with waiver of informed consent. The algorithm developers had no role in the design, analysis, or reporting of this study.

We retrospectively included consecutive patients older than age 18 years between January 2016 and December 2020 who underwent posteroanterior chest radiography at Trillium Health Partners, a high-volume, full service, three-site hospital system that serves the ethnically diverse population of Mississauga, Ontario, Canada.

All chest radiographs were extracted from our picture archiving and communication system. Data necessary for subgroup analysis were extracted, and remaining Digital Imaging and Communications in Medicine (DICOM) metadata elements were removed. Associated radiology reports were also extracted, and regular expression functions were used to de-identify names and dates.

A total of 293 881 posteroanterior chest radiographs were initially included. Duplicate studies were removed, leaving 207 963 chest radiographs. Images that could not be read by one or more algorithms were excluded, leaving a final dataset of 197 540 chest radiographs (Tables 1, 2).

"Ground Truth" Extraction from Radiology Reports

Although multiple human annotations are the reference standard for ground truth generation, they are costly and limit dataset size (19). To perform labeling at scale, we used multiclass natural language processing (NLP) tools to classify local radiology reports into one or more of 14 categories. These categories included the following: *(a)* 12 pathology classes; *(b)* a support device class; and *(c)* a "no finding" class, a catch-all category indicating the absence of any clinically relevant abnormality. Because findings on chest radiographs represent a "long-tailed" distribution with few common findings and many uncommon findings, the "no finding" class was trained by the NLP tool developers to represent 53 additional findings beyond the 12 pathology classes (eg, osteopenia, aortic aneurysm) (20,21). If the "no finding" class was positive, the image was considered to be absent of any abnormality, including the "long tail," and was labeled normal; if the "no finding" class was negative, abnormality was present, and the image was labeled abnormal.

We validated the performance of two open-source NLP classifiers (the CheXpert [1] and CheXbert [22] NLP labelers) on radiology reports at our institution by comparing predicted classifications to manual annotations. In contrast to original studies that used only the report's summary or conclusion, our analysis included the entire report because of variation in report structures across institutions and radiologists. For this analysis, Trillium Health Partners radiology reports were manually classified by two independent investigators (M. Ahluwalia and J.S., both 3rd-year medical students), and conflicts were resolved through consensus. Reports that solely described a lack of interval change, referred to findings in other scans, or could not be interpreted by one or more NLP or image classifiers were excluded. A total of 502 reports were included (similar in number to CheXpert Image's test-set size of 500) (1).

To measure the effect of NLP error on image classification performance, we compared the performance of each image classifier on the 502-image dataset against each type of report label (ie, manual, CheXpert, and CheXbert). Recognizing that radiology reports incorporating information from previous scans may bias NLP algorithm performance, we also determined image classifier performance on chest radiographs that were the first for a patient in the dataset (ie, with no comparison). This analysis was restricted to emergency department and outpatient chest radiographs.

Image Classifier Performance

We performed external testing of three open-source state-of-the-art chest radiograph disease classification models (trained on the largest publicly available chest radiograph datasets: CheXpert Image [1], MIMIC-CXR [17], and Chest X-ray-14 [18]) and one proprietary third-party model by comparing predictions to ground truth measurements determined by the CheXpert NLP labeler on associated radiology reports. These classifiers have been shown to be state-of-the art in chest radiograph classification when compared with other models trained on the same datasets (14). To enable direct comparisons between algorithms, each with different schemas, we collapsed multiclass predictions into a single binary prediction (ie, normal or abnormal) (23). The presence of support devices was not considered an abnormality. The operating thresholds were taken "as-is" from the model

Table 1: Characteristics of the 197 540-Image Dataset Used to Evaluate Chest Radiograph Classifiers

Subgroup	Normal	Abnormal	Total
Overall	101 002 (51)	96 538 (49)	197 540
Patient age			
>65 y	23 231 (31)	54 027 (69)	77 258 (39)
40–65 y	49 532 (59)	33 875 (41)	83 407 (42)
<40 y	28 239 (76)	8636 (24)	36 875 (19)
Patient setting			
Emergency	53 267 (63)	31 213 (37)	84 480 (43)
Inpatient	8301 (32)	17 369 (68)	25 670 (13)
Outpatient	35 741 (46)	42 216 (54)	77 957 (39)
Intensive care unit	840 (28)	2140 (72)	2980 (2)
No location specified	2849 (44)	3595 (56)	6444 (3)
EHR-identified sex			
Male	48 401 (49)	49 760 (51)	98 161 (50)
Female	52 595 (53)	46 774 (47)	99 369 (50)
Name-based ancestry			
Greater European	64 445 (49)	67 535 (51)	131 980 (67)
Greater African/Indian	31 140 (57)	23 198 (43)	54 338 (28)
Greater East Asian	5417 (48)	5805 (52)	11 222 (6)
Chest radiograph modality			
GE type 1	1607 (48)	1775 (52)	3382 (2)
Agfa type 1	2289 (50)	2310 (50)	4599 (2)
GE type 2	18 001 (50)	18 234 (50)	36 235 (18)
Canon type 1	2919 (52)	2717 (48)	5636 (3)
Philips type 1	26 438 (46)	30 644 (54)	57 082 (29)
Carestream type 1	15 106 (54)	13 095 (54)	28 201 (14)
Toshiba type 1	30 164 (56)	23 768 (44)	53 932 (27)
Varian type 1	343 (50)	347 (50)	690 (< 1)
Other modalities	4135 (53)	3648 (47)	7783 (4)
No. of classes			
0 findings	101 002	19 699*	120 701 (61)
1 finding	0	42 582	42 582 (22)
2 findings	0	22 213	22 213 (11)
3 findings	0	8754	8754 (4)
4 findings	0	2601	2601 (1)
5 findings	0	588	588 (<1)
≥6 findings	0	101	101 (<1)

Note.—Data are presented as numbers of images, with percentages in parentheses. EHR = electronic health record.
* Abnormal chest radiographs present in the 0 findings category represent abnormalities that were not captured by the CheXpert natural language processing algorithm's 12 pathology classes.

developers; to permit external testing, the models were not altered or retrained.

To aid in result interpretation, a random classifier that randomly associated a normal or abnormal prediction to each image with a rate equivalent to overall prevalence was used as a reference point (Table 1). This type of "dummy" baseline model is used in the setting of imbalanced classes or subgroups to help readers interpret performance metrics such as accuracy, positive predictive value (PPV), and negative predictive value (NPV), which can be falsely inflated or deflated (eg, a model with 90% accuracy in a population with 10% prevalence of disease does not demonstrate much skill).

Table 2: Description of the Pathology-specific Subgroups of the 197 540-Image Dataset

Subgroup	Normal	Abnormal	Total
Solitary findings			
Enlarged cardiomediastinum	0	6275	6275
Cardiomegaly	0	7609	7609
Lung lesion	0	2732	2732
Lung opacity	0	8674	8674
Edema	0	454	454
Consolidation	0	413	413
Pneumonia	0	1502	1502
Atelectasis	0	5288	5288
Pneumothorax	0	1395	1395
Pleural effusion	0	3537	3537
Pleural other	0	1123	1123
Fracture	0	3580	3580
Support device	2936	0	2936
General findings			
Enlarged cardiomediastinum	0	14 306	14 306
Cardiomegaly	0	17 734	17 734
Lung lesion	0	7652	7652
Lung opacity	0	25 916	25 916
Edema	0	3260	3260
Consolidation	0	3547	3547
Pneumonia	0	6455	6455
Atelectasis	0	16 052	16 052
Pneumothorax	0	3720	3720
Pleural effusion	0	18 301	18 301
Pleural other	0	3517	3517
Fracture	0	6764	6764

Note.—Data are presented as numbers of images.

Subgroup Analysis

To provide a granular understanding of each classifier's performance, subgroup analysis was conducted on patient, setting, and pathology variables.

Patient subgroups included age, electronic health record (EHR)–reported sex, and ancestry. Patient age was divided into the following categories: older than 65 years, 40–65 years, and younger than 40 years. EHR-reported sex was derived from the EHR as male or female; gender information was not stored in the DICOM metadata. Patient ancestry was determined through name-based identification and grouped into three large ancestral groups (greater European, greater African/Indian, greater East Asian) as a proxy for ethnicity (Appendix S1).

Setting subgroups included patient location and chest radiograph modality. Patient location was defined as outpatient, inpatient, intensive care unit (ICU), or emergency department based on the location of the patient at the time of image order. The chest radiograph modality (equipment model) was obtained from the DICOM metadata. Modalities with fewer than 500 chest radiographs were excluded from this analysis.

Pathology subgroups were determined from the CheXpert NLP results. The solitary findings subgroup comprised radiology reports containing exactly one positive pathology class and the absence of all other abnormalities. For the number of classes analysis, a sum of positive pathology classes (up to 12 abnormalities) was taken to represent the number of classes in each chest radiograph. For example, if an image contained a pleural effusion and consolidation, it would be assigned to the two-classes category. Finally, a support devices analysis included radiology reports containing only support devices and no pathology classes to determine the effect of support devices on model specificity. For subgroups with only abnormal chest radiographs, we report only sensitivity; for subgroups with only normal chest radiographs, we report only specificity (Tables 1, 2).

Statistical Analysis

Performance was measured using the metrics of accuracy, precision (PPV), NPV, recall (sensitivity), specificity, F1 score, and Matthews correlation coefficient (MCC). In contrast to traditional performance metrics, the MCC generates a high

value only if a binary predictor accurately classifies both positive and negative cases (penalizing false classifications) and, unlike accuracy, is robust against class imbalance. The MCC is a correlation measure between actual and predicted values (1 represents perfect classification; -1 represents perfect misclassification) (24,25). The formula for each metric can be found in Appendix S1. Bootstrap resampling (10 000 replicates, with each replicate containing the same number of images as the group being analyzed) was used to determine 95% CIs.

To measure the effect of NLP error on image classification performance, we compared the performance of each image classifier on the 502-image dataset against each type of report label (ie, manual, CheXpert, and CheXbert). Using bootstrap resampling (502 samples per replicate, 10 000 replicates), we generated 95% CIs for each performance metric. Empirical *P* values were estimated from 95% CIs using the formula $P = (r+1)/(n+1)$, where n represents the number of bootstrap simulations and r represents the number of simulations that produce a value higher than that observed in the real data (26). This analysis was conducted in Python software, version 3.9.5.

Table 3: Performance of the CheXpert and CheXbert Natural Language Processing Algorithms Compared with Manual Labels in the 502-Chest Radiograph Dataset

Overall	CheXpert	CheXbert
Accuracy	473/502 (94)	465/502 (93)
Precision/PPV	226/250 (90)	228/262 (87)
NPV	247/252 (98)	237/240 (99)
Recall/sensitivity	226/231 (98)	228/231 (99)
Specificity	247/271 (91)	237/271 (87)
F1 score	0.94	0.92
MCC	0.89	0.86

Note.—Unless otherwise noted, data are presented as proportions of images, with percentages in parentheses. MCC = Matthews correlation coefficient, NPV = negative predictive value, PPV = positive predictive value.

Results

NLP Algorithm Performance

The NLP validation dataset (n = 502) is described in Table S1. Table 3 describes the performance of the CheXpert and CheXbert NLP algorithms compared with manual labels.

In the NLP validation dataset, the CheXpert NLP demonstrated an accuracy of 94% (473 of 502), which is similar to the agreement between multiple radiologist annotators (23,27,28). In addition, we found no evidence of a difference in third-party classifier performance for all metrics, except for NPV, between manual labels and CheXpert NLP-generated labels (Fig 1; Table S2). Performance of image classifiers on the first chest radiograph of patients was lower, indicating that reports referring to previous studies was not a detriment to NLP labeler performance (Fig S1). CheXpert outperformed CheXbert and was selected to generate ground truth labels for the 197 540-image dataset.

Chest Radiograph Classifier Performance

Data characteristics.— The Trillium Health Partners dataset comprised 197 540 posteroanterior chest radiographs, with 48.9% (n = 96 538) of scans labeled abnormal by the CheXpert NLP algorithm. Patient age ranged from 18 to older than 105 years. We found equal proportions of male and female patients, skewing toward greater European ancestry, emergency department patients, and outpatients. The full dataset description is shown in Tables 1 and 2.

Overall results.— The third-party algorithm demonstrated the highest overall performance (accuracy of 77% [152 500 of 197 540], MCC of 0.55), followed by algorithms trained on the CheXpert Image, MIMIC-CXR, and Chest X-ray-14 datasets

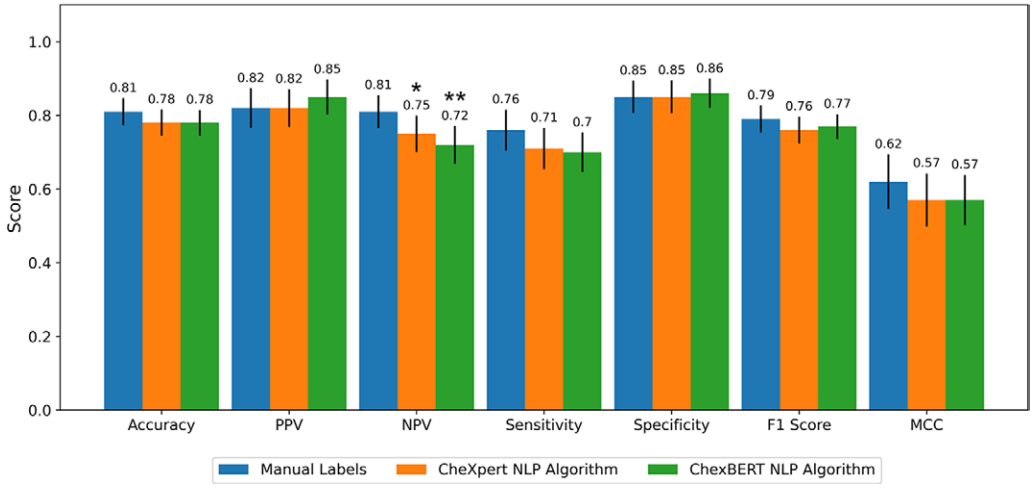

Figure 1: Performance of the third-party algorithm on the 502-image dataset with use of manual labels versus natural language processing–generated labels as ground truth. Error bars show 95% CIs. *P < .05 compared with manual labels, **P < .01 compared with manual labels. Full statistical results shown in Table S2. BERT = bidirectional encoder representations from transformers, MCC = Matthews correlation coefficient, NPV = negative predictive value, PPV = positive predictive value.

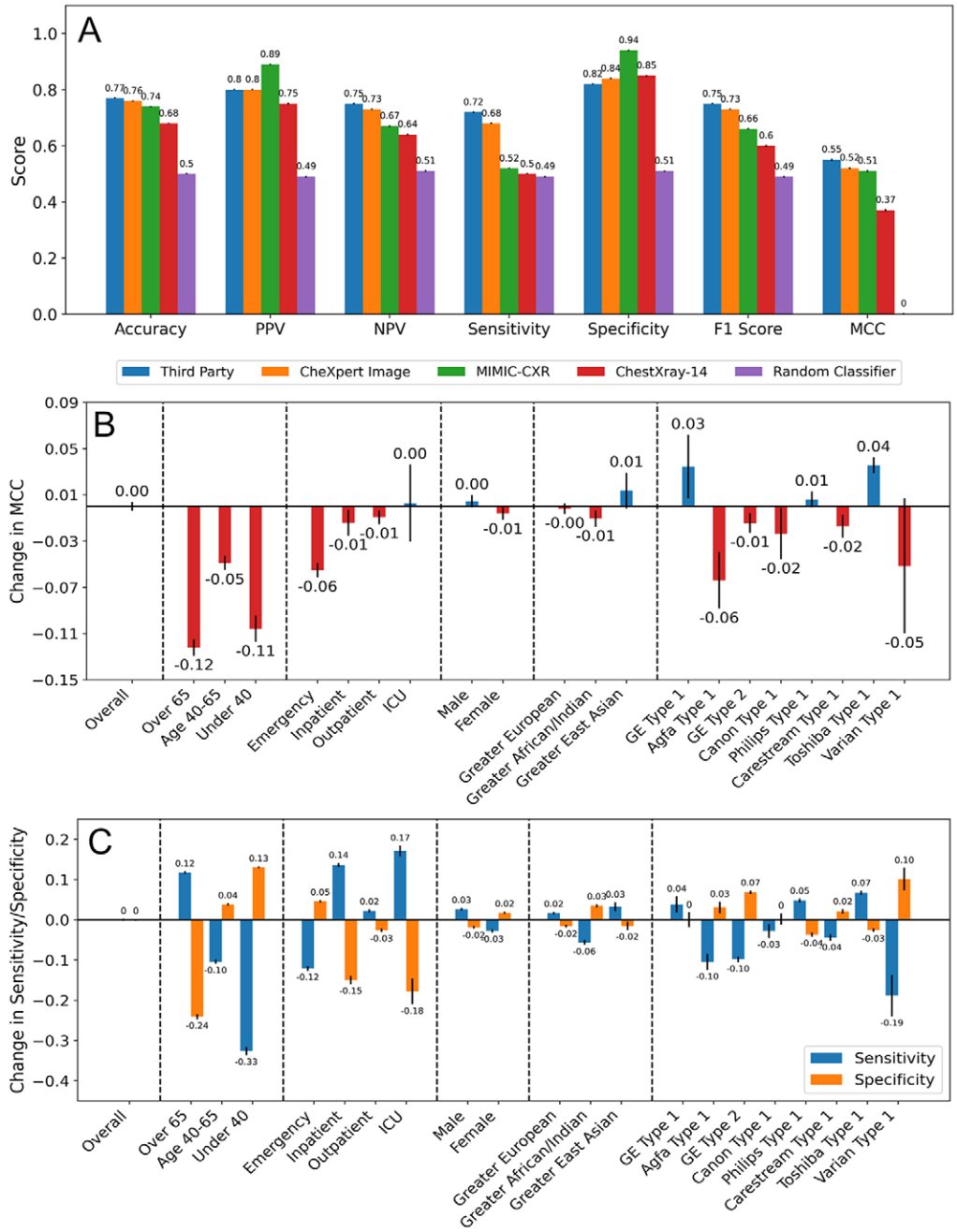

Figure 2: **(A)** Overall performance of the four deep learning classifiers and the random classifier on the entire 197 540-image dataset, with error bars showing 95% CIs. **(B)** Change in Matthews correlation coefficient (MCC) of the third-party algorithm for the subgroups of patient age, patient setting, patient electronic health record–identified sex, patient name–generated ancestry, and chest radiography modality compared with the overall dataset, with error bars showing 95% CIs. **(C)** Change in sensitivity and specificity of the third-party algorithm for each subgroup compared with the overall dataset, with error bars showing 95% CIs. ICU = intensive care unit, NPV = negative predictive value, PPV = positive predictive value.

(Fig 2A). All image classifiers showed a similar pattern of lower sensitivity and higher specificity and lower NPV and higher PPV. Figure 2B and 2C describe differences in MCC, sensitivity, and specificity between the overall dataset and individual subgroups. Results for each subgroup can be found in Figures S2–S8.

Patient subgroups.— All algorithms demonstrated markedly lower sensitivity and higher specificity when analyzing chest radiographs of patients younger than 40 years, with the opposite results for patients older than 65 years (Fig 2; Figs S5, S6). Only 24% of studies were labeled abnormal in the subgroup younger than 40 years, compared with 69% of those in the subgroup older than 65 years. Abnormal chest radiographs of patients younger than 40 years compared with those of patients older than 65 years contained fewer pathology classes (mean, 1.40 vs 1.76, respectively), suggesting that those younger than 40 years had, as expected, fewer abnormal findings (Tables 1, 2; Fig S9).

The algorithms demonstrated similar performance on EHR-identified sex and name-derived ancestry subgroups (Figs S2–S8).

Setting subgroups.— Classifiers demonstrated lower sensitivity and higher specificity on emergency department patients, whereas the opposite was observed for inpatients and ICU patients (Fig 2; Figs S5, S6). Abnormal chest radiographs from the emergency department had fewer pathology classes on average compared with inpatients (mean, 1.51 vs 1.84), suggesting that patients presenting to the emergency department had, as expected, fewer abnormal findings. This variation is not fully explained by differences in age because patients in the same age group demonstrated the same pattern across settings (ie, intersectionality) (Fig S10).

Analyzing performance across different chest radiograph modalities showed variation (for the third-party algorithm, sensitivity was 53%–79% [184 of 347; 18 673 of 23 768]; specificity was 79%–92% [20 780 of 26 438; 317 of 343]; and MCC was 0.48–0.58) (Figs S5, S6, S8). The location and case types for each modality were not investigated and may confound and/or explain the variation.

Pathology subgroups.— Classifiers demonstrated lower sensitivity in the categories of enlarged cardiomediastinum, cardiomegaly, lung lesions, pneumonia, pneumothorax, and fractures. In fact, the random classifier outperformed all image classifiers in detecting a solitary enlarged cardiomediastinum, pneumonia, and fracture (Fig S11). However, the classifiers showed high sensitivity in detecting pleural effusions and consolidations.

Classifier performance was universally lower in one clinically important scenario: cases with a solitary finding. The third-party algorithm correctly classified 65% (27 595 of 42 582) of images with solitary findings as abnormal versus 97% (11 738 of 12 044) of images with three, four, five, or six or more finding classes. Sensitivity decreased for all pathology classes when images with a solitary finding were analyzed (Fig 3, Fig S11).

Classifiers showed 9%–24% decreased specificity on images with support devices but no pathology classes (eg, a central line in an otherwise normal chest radiograph). For example, the third-party algorithm had a false-positive rate of 41% on normal images with support devices (specificity of 59% [1797 of 2936]) compared with 17% on normal images without support devices (specificity of 83% [81 387 of 98 066]) (Fig S12).

Operating curves.— Comparing performance for different subgroups can be facilitated by plotting subgroup operating points on a receiver operating characteristic curve (Fig 4). Viewed this way, some subgroup operating points represent trade-offs along an operating curve (eg, male vs female: the algorithm is more sensitive but less specific in male patients), whereas other operating points represent increased or decreased performance, as located on shifted curves (eg, age >65 years vs age 40–65 years, emergency department vs outpatient).

Discussion

The generalization gap exposed during external testing of machine learning classifiers is well established (12); our study extends the understanding of these performance gaps to critical patient, setting, and pathology subgroups. Although chest radiograph classifiers performed with accuracy nearing 80%, large-scale external testing demonstrates that performance is not uniform, varying widely by patient and disease context. We review important subgroup-specific generalization gaps below, but we first briefly discuss how our methods validated a scalable labeling approach for external testing.

We demonstrated that NLP algorithms accurately label normal and abnormal chest radiograph radiology reports in external testing experiments, even when the entire report is included, which can lower associated costs and preprocessing requirements. Open-source NLP tools correctly classified more than 93% of radiology reports compared with human labels; across measures (except NPV), we found no evidence of a difference in performance of image classifiers on manual versus NLP labels. Most (75%) NLP algorithm false-positive results occurred in the category of enlarged cardiomediastinum and cardiomegaly. Errors in NLP algorithm performance included detecting the indication for the chest radiograph as an abnormality, misinterpreting terms used to describe the cardiomediastinum, and missing more nuanced negations (eg, "pneumothorax seen previously is no longer present"). Addressing these errors would further improve the accuracy of NLP labeling tools.

Regarding subgroup-specific generalization gaps, we demonstrated an overall pattern that suggests increased performance in detecting grossly abnormal chest radiographs over solitary findings. This is demonstrated by decreased performance at classifying images with solitary findings (sensitivity ≤ 65%, an absolute 27% decrease compared with two or more pathology classes for the third-party algorithm) and increased performance on chest radiographs with three or more pathology classes (sensitivity ≥ 97%). The number of classes positively correlates with age and more acute care settings; older patients, inpatients, and ICU patients are more likely to have grossly abnormal chest radiographs, which may explain the increased sensitivity and PPV in these subgroups and decreased sensitivity and PPV in young and emergency department patients. This hidden stratification has concerning implications for algorithm deployment because radiologists would benefit most from decision support in challenging solitary finding cases (eg, a subtle pneumothorax); instead, classifiers provide reliable support only when multiple findings are present. In addition, low sensitivity in solitary findings should raise alarm when users consider deploying these algorithms as a screening tool to rule out important abnormalities or deprioritize chest radiographs in radiology workflows.

Exploring protected attribute fairness (eg, sex, ancestry), three of the classifiers demonstrated lower sensitivity on female patients compared with male patients (0%–7%), and all classifiers demonstrated lower sensitivity in those with names of African/Indian ancestry compared with other ethnic groups (3%–10%). This decrease in sensitivity could be interpreted as "underdiagnosis." Seyyed-Kalantari et al (15) previously demonstrated that state-of-the-art classifiers trained on public datasets showed bias

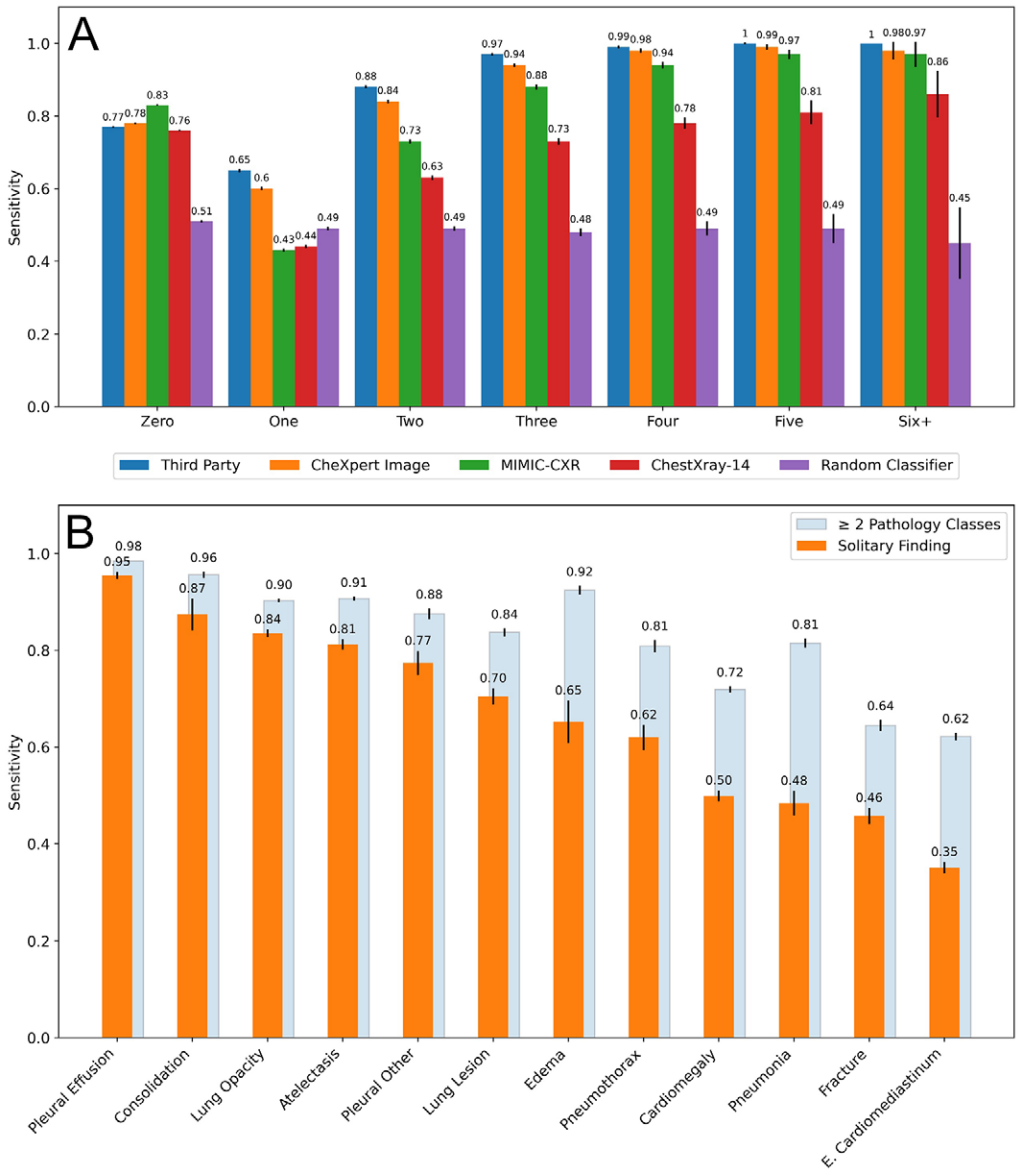

Figure 3: **(A)** Performance of the four deep learning classifiers stratified by the number of pathology classes present, noting high sensitivity (third party classifier, ≥ 97%) in detecting abnormal chest radiographs with three or more findings and low sensitivity (all classifiers, ≤ 65%) in detecting abnormal chest radiographs with solitary findings. Error bars show 95% CIs. **(B)** Performance of the third-party classifier in detecting abnormal chest radiographs with a solitary finding compared with detecting abnormal chest radiographs with two or more pathology classes. Graph shows decreased performance on solitary findings and performance that is dependent on the finding type present. Error bars show 95% CIs. E = enlarged.

against patients who were female or Hispanic. In another study, Seyyed-Kalantari et al (14) found that these classifiers underdiagnosed underserved populations at a rate exacerbated by intersectional identities. Here, performance differences in sex and ancestry subgroups may represent movements along a receiver operating characteristic curve as opposed to decreased overall performance, as the MCC between groups does not vary substantially and decreased sensitivity is compensated by increased specificity. Unfortunately, few studies analyze the effect of protected attributes on radiology classifier performance despite it being critical in preventing the propagation of health disparities.

In addition, decreased specificity on normal images with support devices suggests algorithms may erroneously label a "normal" chest radiograph as "abnormal" because a support device is present, even in the absence of true cardiothoracic abnormality. These "spurious correlates" are especially relevant in inpatient and ICU populations and contribute to hidden stratification (29). Concordant with our findings, Chen et al (30) described the presence of support devices as a statistically significant predictor of misclassification by image classifiers. Future work is required to avoid device-induced misclassification and improve classifier robustness.

We believe subgroup analysis during external testing is critical to informing safe and equitable deployment of artificial intelligence into clinical practice. For example, a triage tool in the emergency department demands high performance for patients

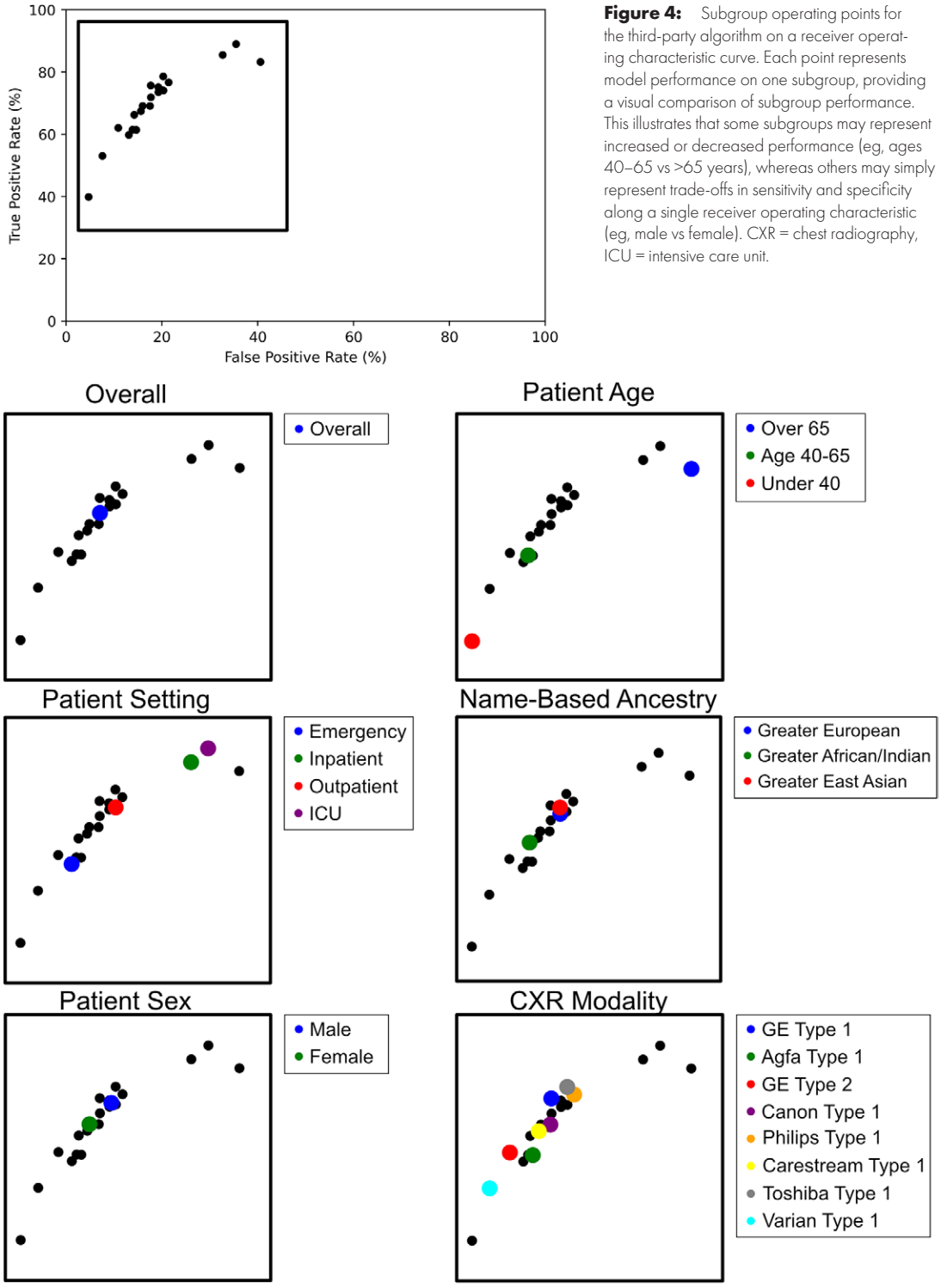

Figure 4: Subgroup operating points for the third-party algorithm on a receiver operating characteristic curve. Each point represents model performance on one subgroup, providing a visual comparison of subgroup performance. This illustrates that some subgroups may represent increased or decreased performance (eg, ages 40–65 vs >65 years), whereas others may simply represent trade-offs in sensitivity and specificity along a single receiver operating characteristic (eg, male vs female). CXR = chest radiography, ICU = intensive care unit.

in the emergency department, younger patients, and emergent findings, such as a pneumothorax. However, all classifiers studied performed poorly on younger, emergency department patients as well as solitary pneumothoraces. We see consistent performance disparities across classifiers despite each being pretrained on large datasets gathered from different institutions. Thus, such outcomes are not random, and health systems deploying such classifiers must decide whether they are safe to deploy and, if so, what restrictions, training, or warnings must accompany deployment. Sendak et al (31) describes a risk communication tool, the "Model Facts" label, which succinctly describes actionable, clinically relevant information to inform users about "how, when, how not, and when not to incorporate model output into clinical decisions." To create such a tool, subgroup analysis is essential. In

this case, a warning label might alert clinicians to poor sensitivity on chest radiographs with certain solitary findings while also providing information to developers and policymakers on the shortcomings of these classifiers. Finally, subgroup performance should be intentionally monitored for drifts over time to ensure ongoing quality, safety, and equity (32).

Our analysis had limitations. First, using a single radiology report as a reference standard is a trade-off that values scalability over certainty: Radiologists do not always report every finding, and one radiologist's report may not be concordant with another's (33,34). In addition, the CheXpert NLP algorithm introduces error; however, it was validated in this study and its error lies within statistical uncertainty of multiple radiologist annotators (27,28). Second, specific annotation schemas exclude many clinically relevant conditions, such as pneumomediastinum. There is also class overlap between finding categories, such as consolidation and pneumonia, which counted as two classes in the analysis of number of classes. Each classifier was trained using different annotation schemas; a negative output from one did not reflect the same excluded abnormalities as another (23). Third, there may also be confounding variables between different subgroups, and this was not investigated beyond age and setting subgroups. Fourth, only posteroanterior chest radiographs were used because of classifier limitations, and our results are not generalizable to other chest radiograph views. Fifth, because this was a single-center study, the numeric results and specific gaps may also not be generalizable to other institutions; the precise generalization gaps across subgroups will undoubtedly vary at other institutions with different populations. Finally, name-based patient ancestry is an imperfect method of determining ethnicity-based subgroups but was the most practical method, given the lack of routine collection of self-identified ethnic and racial identifiers at Trillium Health Partners and the necessity of conducting such an analysis.

In conclusion, external testing of four chest radiograph classifiers demonstrated consistent differences in performance on patient, setting, and pathology subgroups. Subgroup analysis is critical when validating image classifiers to identify risks to safe and equitable deployment.

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

Data sharing: Data generated or analyzed during the study are available from the corresponding author by request.

Disclosures of conflicts of interest: M. Ahluwalia Digital Supercluster Canada is funder of this research project (no role in any research activities). **M. Abdalla** Vanier Canada Graduate Scholarship funding (paid directly to author). **J.S.** No relevant relationships. **L.S.K.** No relevant relationships. **M.H.** Support from Trilium Health Partners; consulting fees from Trilium Health Partners. **A.A.** No relevant relationships. **B.F.** Support from Trillium Health Partners Foundation and Digital Supercluster Canada; leadership role on HaloHealth Angel Network Board; stocks or stock options in PocketHealth, Phelix, and Eva Medical.

References

1. 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(01):590–597.
2. NHS. Diagnostic imaging dataset annual statistical release 2019/20. London, UK: Department of Health, 2020.
3. Crisp N, Chen L. Global supply of health professionals. N Engl J Med 2014;370(10):950–957.
4. Rimmer A. Radiologist shortage leaves patient care at risk, warns royal college. BMJ 2017;359:j4683.
5. Park SH, Han K. Methodologic guide for evaluating clinical performance and effect of artificial intelligence technology for medical diagnosis and prediction. Radiology 2018;286(3):800–809.
6. Whitman GJ, Moseley TW. Stand-alone machine learning: more work is needed. Radiology 2022;302(1):105–106.
7. Ting DSW, Cheung CYL, Lim G, et al. Development and validation of a deep learning system for diabetic retinopathy and related eye diseases using retinal images from multiethnic populations with diabetes. JAMA 2017;318(22):2211–2223.
8. Thian YL, Ng D, Hallinan JTPD, et al. Deep learning systems for pneumothorax detection on chest radiographs: a multicenter external validation study. Radiol Artif Intell 2021;3(4):e200190.
9. Nam JG, Park S, Hwang EJ, et al. Development and validation of deep learning–based automatic detection algorithm for malignant pulmonary nodules on chest radiographs. Radiology 2019;290(1):218–228.
10. Kitamura G, Deible C. Retraining an open-source pneumothorax detecting machine learning algorithm for improved performance to medical images. Clin Imaging 2020;61:15–19.
11. Keskar NS, Mudigere D, Nocedal J, Smelyanskiy M, Tang PTP. On large-batch training for deep learning: generalization gap and sharp minima. arXiv 1609.04836 [preprint] https://arxiv.org/abs/1609.04836. Posted September 15, 2016. Accessed March 2023.
12. Yu AC, Mohajer B, Eng J. External validation of deep learning algorithms for radiologic diagnosis: a systematic review. Radiol Artif Intell 2022;4(3):e210064.
13. Riley RD, Ensor J, Snell KIE, et al. External validation of clinical prediction models using big datasets from e-health records or IPD meta-analysis: opportunities and challenges. BMJ 2016;353:i3140. [Published correction appears in BMJ 2019;365:l4379.]
14. Seyyed-Kalantari L, Zhang H, McDermott MBA, Chen IY, Ghassemi M. Underdiagnosis bias of artificial intelligence algorithms applied to chest radiographs in under-served patient populations. Nat Med 2021;27(12):2176–2182.
15. Seyyed-Kalantari L, Liu G, McDermott M, Chen IY, Ghassemi M. CheXclusion: fairness gaps in deep chest X-ray classifiers. Pac Symp Biocomput 2021;26:232–243.
16. Char DS, Shah NH, Magnus D. Implementing machine learning in health care - addressing ethical challenges. N Engl J Med 2018;378(11):981–983.
17. Johnson AEW, Pollard TJ, Berkowitz SJ, et al. MIMIC-CXR, a de-identified publicly available database of chest radiographs with free-text reports. Sci Data 2019;6(1):317.
18. 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. arXiv 1705.02315 [preprint] https://arxiv.org/abs/1705.02315. Posted May 5, 2017. Accessed March 2023.
19. Willemink MJ, Koszek WA, Hardell C, et al. Preparing medical imaging data for machine learning. Radiology 2020;295(1):4–15.
20. Holste G, Wang S, Jiang Z, et al. Long-tailed classification of thorax diseases on chest x-ray: a new benchmark study. arXiv 2208.13365 [preprint] https://arxiv.org/abs/2208.13365. Posted August 29, 2022. Accessed March 2023.
21. Zhang Y, Kang B, Hooi B, Yan S, Feng J. Deep long-tailed learning: a survey. arXiv 2110.04596 [preprint] https://arxiv.org/abs/2110.04596. Posted October 9, 2021. Accessed March 2023.
22. Smit A, Jain S, Rajpurkar P, Pareek A, Ng AY, Lungren MP. CheXbert: combining automatic labelers and expert annotations for accurate radiology report labeling using BERT. arXiv 2004.09167 [preprint] https://arxiv.org/abs/2004.09167. Posted April 20, 2020. Accessed March 2023.
23. Abdalla M, Fine B. Hurdles to artificial intelligence deployment: noise in schemas and "gold" labels. Radiol Artif Intell 2023;5(2):e220056.
24. Chicco D, Jurman G. The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation. BMC Genomics 2020;21(1):6.
25. Chicco D, Jurman G. The Matthews correlation coefficient (MCC) should replace the ROC AUC as the standard metric for assessing binary classification. BioData Min 2023;16(1):4.

26. North BV, Curtis D, Sham PC. A note on the calculation of empirical P values from Monte Carlo procedures. Am J Hum Genet 2002;71(2):439–441.
27. Fowler KJ, Tang A, Santillan C, et al. Interreader reliability of LI-RADS version 2014 algorithm and imaging features for diagnosis of hepatocellular carcinoma: a large international multireader study. Radiology 2018;286(1):173–185.
28. Chung R, Rosenkrantz AB, Bennett GL, et al. Interreader concordance of the TI-RADS: impact of radiologist experience. AJR Am J Roentgenol 2020;214(5):1152–1157.
29. Oakden-Rayner L, Dunnmon J, Carneiro G, Ré C. Hidden stratification causes clinically meaningful failures in machine learning for medical imaging. arXiv 1909.12475 [preprint] https://arxiv.org/abs/1909.12475. Posted September 27, 2019. Accessed March 2023.
30. Chen E, Kim A, Krishnan R, Long J, Ng AY, Rajpurkar P. CheXbreak: misclassification identification for deep learning models interpreting chest x-rays. arXiv 2103.09957 [preprint] https://arxiv.org/abs/2103.09957. Posted March 18, 2021. Accessed March 2023.
31. Sendak MP, Gao M, Brajer N, Balu S. Presenting machine learning model information to clinical end users with model facts labels. NPJ Digit Med 2020;3(1):41.
32. Oakden-Rayner L, Gale W, Bonham TA, et al. Validation and algorithmic audit of a deep learning system for the detection of proximal femoral fractures in patients in the emergency department: a diagnostic accuracy study. Lancet Digit Health 2022;4(5):e351–e358.
33. Peng JM, Qian CY, Yu XY, et al. Does training improve diagnostic accuracy and inter-rater agreement in applying the Berlin radiographic definition of acute respiratory distress syndrome? A multicenter prospective study. Crit Care 2017;21(1):12.
34. Moncada DC, Rueda ZV, Macías A, Suárez T, Ortega H, Vélez LA. Reading and interpretation of chest X-ray in adults with community-acquired pneumonia. Braz J Infect Dis 2011;15(6):540–546.