MELBA Journal Machine Learning for Biomedical Imaging

Exploring Fairness and Performance Drivers Across State-of-the-Art Pulmonary Nodule Detection Algorithms

John McCabe [1], Daryl O. Cheng [1], Andrew Crossingham [4], Junaid Choudhary [4], Samantha L Quaife [7], Tanya Patrick [2], Monica Mullin [2,5], Amyn Bhamani [2], Esther Arthur-Darkwa [6], Aoife Walker [6], Arjun Nair [4], Alan Hackshaw [6], SUMMIT consortium, Sam M. Janes* [2], Joseph Jacob* [1], Carole H. Sudre* [8,3] **These authors contributed equally as joint last authors.

1 Satsuma Lab, Hawkes Institute, University College London, United Kingdom
2 Lungs for Living Research Centre, UCL Respiratory, University College London, United Kingdom
3 Hawkes Institute, University College London, United Kingdom
4 University College London Hospitals NHS Foundation Trust, London, United Kingdom
5 Department of Respirology, University of British Columbia, Vancouver, Canada
6 CRUK & UCL Cancer Trials Centre, University College London, London, United Kingdom
7 Centre for Cancer Screening, Prevention, Detection and Early Diagnosis, Wolfson Institute of Population Health, Barts and The London School of Medicine and Dentistry, Queen Mary University of London, London, United Kingdom
8 Unit for Lifelong Health and Ageing at UCL, Department of Population Science and Experimental Medicine, University College London, United Kingdom

Abstract

Lung cancer is the leading cause of cancer-related deaths in the UK. Its high mortality rate is primarily due to its asymptomatic nature in the early stages, leading to late-stage diagnoses. However, effective early detection methods, such as Low-Dose Computed Tomography (LDCT), and treatments for early-stage disease make lung cancer an ideal candidate for screening. The UK Government aims to implement a national lung cancer screening programme targeting high-risk populations by 2029. This will significantly increase the workload on an already stretched radiology workforce, driving the adoption of computer-aided detection (CADe) systems to support radiologists. The datasets used to train these algorithms are typically drawn from previous lung cancer screening trials and studies (National Lung Screening Trial Research Team (2011); de Koning et al. (2020)), which often lack balanced representation of protected groups, such as sex and ethnicity. This project examines whether training nodule detection algorithms on low-dose computed tomography (LDCT) scans from a London-based lung screening study, where these groups are typically under-represented, affects algorithm performance for under-represented categories. Our results indicate that overall performance remains equitable across all categories, even when trained on unbalanced datasets. The discriminative performance of deep learning-based pulmonary nodule detection algorithms is primarily driven by the composition of the dataset, specifically, the relative proportion of nodule types and sizes, rather than by protected attributes such as sex or ethnic group. The features learned from the nodules themselves drive detection outcomes, meaning that in populations where the prevalent nodule characteristics closely match the training data, performance is likely to be strong. While this study found no demographic disparities for nodule detection, there is no guarantee that this will be true across all populations, particularly those in populations where cancer risk predominates within different nodule distributions. This study provides an early assessment of performance variations of deep learning models across under-represented groups within a standard lung cancer screening dataset. While previous research has focused on improving how well nodule detection algorithms identify pulmonary nodules, this study uniquely focuses on demographic performance disparities and the impact of training data composition and algorithm design on model generalisability. The findings highlight critical considerations for the deployment of CADe systems in lung cancer screening, ensuring equitable performance across diverse patient populations. Our code is available at `https://github.com/johnmccabe44/fairness-in-nodule-detection`.

Keywords

Nodule Detection Algorithms, Fairness in AI, Lung Cancer Screening

Article informations

https://doi.org/https://doi.org/10.59275/j.melba.2025-6838

Received: 2025-04, Published 2025-12

Corresponding author: j.mccabe@ucl.ac.uk
Special issue: Special issue on Fairness of AI in Medical Imaging (FAIMI) 665
Guest editors: Veronika Cheplygina, Aasa Feragen, Andrew King, Ben Glocker, Enzo Ferrante, Eike Petersen, Esther Puyol-Antón, Melanie Ganz-Benjaminsen

1. Introduction

Lung cancer is the leading cause of cancer death and most diagnosed cancer worldwide (Bray et al. (2024)). It is responsible for the most cancer deaths in the UK, with on average 35,000 deaths per year (Cancer Research UK (2015)). This corresponds to 20% of all cancer deaths. The main reason for the high mortality rate is that lung cancer is often asymptomatic in the early stages, and therefore tends to be diagnosed when the cancer is at an advanced clinical stage. If lung cancer is diagnosed early, treatments are more likely to be effective, and the probability of survival improves. Patients diagnosed at early stages (I or II) of lung cancer have a 5-year survival rate ranging from 53% to 92%, while patients at later stages (III and IV) have 5-year survival rates between 10% and 36% (Cancer Research UK (2015)).

Two major Randomized Control Trials (RCTs) (National Lung Screening Trial Research Team (2011); de Koning et al. (2020)) have demonstrated that Low Dose Computed Tomography (LDCT) is effective in detecting early-stage lung cancer in high-risk populations, leading to a 20-24% reduction in lung cancer mortality. The primary goal of LDCT screening is to identify lung nodules, which are abnormal growths, some of which may represent early-stage cancer. However, detecting malignant nodules is challenging, as most are benign and do not require further evaluation, and some normal lung structures can mimic their appearance. Therefore, a key aspect of nodule assessment is accurately identifying those with a high likelihood of malignancy, referred to as actionable nodules, as they require clinical follow-up. This follow-up can range from immediate referral to a Multi-Disciplinary Team (MDT) for highly suspicious nodules, to scheduled follow-up scans to monitor for growth. To ensure consistency in nodule classification and reporting, the Fleischner Society has established standardized terminology and provides guidelines for follow-up and management (Bankier et al. (2024)).

Beyond distinguishing malignant from benign, nodules are further characterized by their radiological appearance into subtypes, which plays a critical role in clinical decision-making. To illustrate this variability, Figure 1 presents representative examples of solid, part-solid, and non-solid nodules. These examples highlight the wide range of appearances across subtypes, underscoring the difficulty of consistent annotation and the potential for inter-observer variability.

The success of the RCTs has led to both the European Commission and UK National Screening Committee reviewing recommendations for screening of high-risk populations such as current and former smokers. The European Commission has recommended lung cancer screening as part of its *Europe's beating Cancer Plan* (Gyrard et al. (2025))

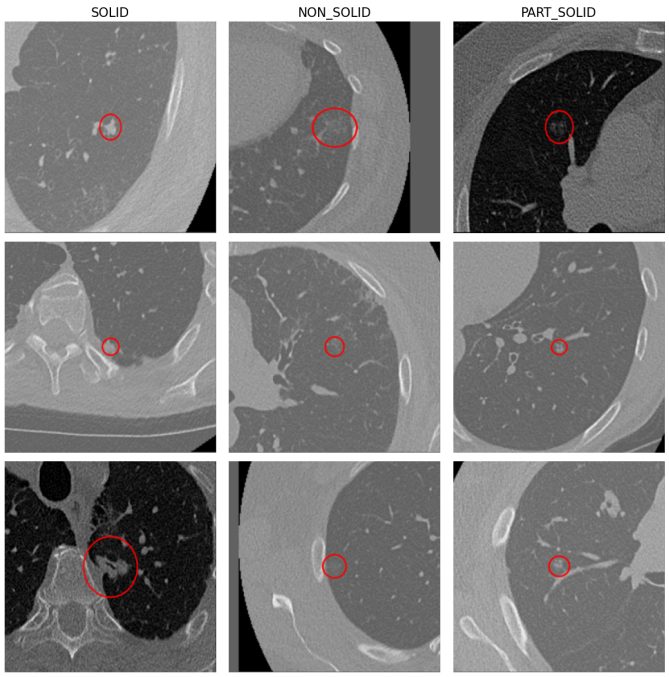

Figure 1: Representative examples of lung nodule types observed in LDCT screening: (left) solid, (middle) part-solid, and (right) non-solid. These images highlight the variability in nodule appearance that contributes to annotation challenges.

with several countries already having started implementation of lung cancer screening programmes (Arnold (2024)). The UK National Screening Committee has recommended the phased roll out of lung cancer screening for high-risk patients across all four UK nations.

The expansion of lung cancer screening programmes, combined with aging populations, will significantly increase the number of LDCT scans requiring reporting by radiologists. However, the widespread shortage of radiologists across many countries (Mirak et al. (2025); The Royal College of Radiologists (2021)), coupled with rising demand, is further exacerbated by factors such as burnout (Parikh et al. (2020)) and an aging workforce approaching retirement. These challenges are likely to place additional strain on existing radiologists, leading to reporting delays, an overburdened workforce, and potential negative impacts on patient care, clinical standards, and overall health outcomes.

Steps are being taken to address these resource challenges, with some countries adopting teleradiology to help address geographical and temporal challenges (Hanna et al. (2020)). Technological advances are another way to respond to these workforce challenges, with artificial intelligence and machine learning (AI/ML) being viewed as a potential solution. AI is gaining significant attention across various sectors. In particular, deep learning (DL), a key subset of AI, has already shown significant promise in healthcare settings, having been successfully deployed in areas such as computer-

aided detection (CADe) for example Veolity (MeVis Medical Solutions AG), Veye Lung Nodules (Aidence), and AI-Rad Companion (Siemens Healthineers), drug discovery (Yang et al. (2019)), and robotic surgery (Beyaz (2020)).

Generalisability and bias continues to be a significant hurdle in DL model development (Buolamwini and Gebru (2018); Wang and Deng (2019); Brandao (2019)), as it depends on training with sufficient data to accurately learn the underlying data distribution. When the training data fails to fully represent the diversity of real-world cases, models struggle to generalise effectively, often resulting in bias. This issue is particularly critical in medical imaging, where biases have been observed due to training on limited datasets, spanning various medical domains and diseases (Puyol-Anton et al. (2021); Weng et al. (2023); Burlina et al. (2021)).

Ensuring fairness when developing AI systems for lung cancer screening is challenging given existing disparities in lung cancer outcomes across sex and ethnic groups. These disparities arise from complex interlocking factors, including socioeconomic deprivation, which is linked to smoking, as well as cultural differences in smoking habits and access to healthcare. If AI systems integrated into lung cancer screening workflows introduce bias, they risk exacerbating these structural inequalities, further widening health disparities.

Several commercial CADe systems have been licensed as second readers for lung cancer screening, aiding radiologists in scan interpretation. To protect intellectual property, the specifics of their model design and training data are deliberately opaque. The datasets available for training these systems are limited, primarily sourced from past studies or publicly available repositories. Widely used datasets like NLST and NELSON also lack diversity. For instance, NLST had more than 90% of participants from White ethnic backgrounds and the NELSON study had a 5:1 male-to-female ratio. Additionally, most of these data sets were collected more than a decade ago. Since then advances in scanner technology have altered scan characteristics, such as resolution and spacing while guidelines on lung nodule taxonomy have evolved, thereby introducing temporal bias in these datasets.

Most academic research on nodule detection algorithms (Setio et al. (2017); Kaggle (2017)), treat all nodules as a single entity, without distinguishing between different nodule types. This is partly due to the limitations of publicly available datasets, which either lack nodule type, annotations or contain too few cases of certain types for meaningful analysis. This contrasts with the fact that distinct types of lung cancer are more likely associated with certain nodule types (Roy et al. (2022)) and consequently have both varying management plans and mortality rates. Furthermore, there is good evidence that the incidence of different lung cancers and nodule types varies across ethnic groups and sex (Chen et al. (2025)).

This paper examines variations in performance across protected characteristics (sex and ethnicity) using multiple state-of-the-art nodule detection algorithms. It investigates whether performance disparities arise from training on datasets where underperforming groups are underrepresented. Additionally, it explores key factors driving model performance.

Our earlier work, presented at the MICCAI 2024 FAIMI Workshop (McCabe et al. (2024)), investigated demographic fairness in lung nodule detection using two state-of-the-art CNN-based algorithms. That study evaluated model performance across sex and ethnic groups at the nodule level, using Free-response Receiver Operating Characteristic (FROC) and Competition Performance Metric (CPM) analyses. We reported initial evidence of performance disparities linked to the composition of training data, particularly the distribution of nodule subtypes.

Building on this foundation, the present work substantially extends our analyses in several directions:

- **Expanded model evaluation:** In addition to CNN based detection algorithms, we evaluate a recently published nodule detection model that integrates transformer blocks. The inclusion of a transformer based model allows us to examine whether the capacity of transformers to capture long-range spatial dependencies and global context improves nodule localisation relative to conventional patch-based CNN approaches, and whether such architectural differences influence fairness across demographic groups.

- **Broader performance metrics:** We extend beyond traditional nodule-level metrics by including mean average precision (mAP). The standard Free Receiver Operating Curve (FROC) metric only checks if the predicted centroid falls within the annotated nodule's radius, ignoring predicted diameter. By adding mAP, we explicitly evaluate the accuracy of both nodule localization and diameter prediction, providing a more comprehensive assessment. We also perform statistical analyses to assess significant differences in performance between demographic groups.

- **Fairness-focused scan-level analysis:** We evaluate four scan-level fairness metrics, true positive rate (TPR), false positive rate (FPR), positive predictive value (PPV), and prevalence ratio (PR), to better capture downstream clinical implications of demographic performance disparities.

- **External validation:** We introduce the Lung Screen Uptake Trial (LSUT) dataset as an external test set, demonstrating the generalisability of our findings across independent screening studies.

- **Calibration assessment:** We assess calibration metrics to evaluate the reliability of model outputs, an important consideration for clinical deployment.

Together, these contributions strengthen the analysis by linking methodological innovation to fairness assessment, and by highlighting both technical and clinical implications of demographic performance disparities in lung cancer screening AI systems. In this work, we adopt a group fairness perspective, defining fairness as the absence of systematic performance differences across demographic subgroups (sex and ethnicity). We operationalise this definition through established detection metrics at both the nodule level (FROC sensitivity, CPM, mAP) and the scan level (TPR, FPR, PPV, PR). Comparable performance across subgroups on these metrics is interpreted as evidence of algorithmic fairness, while systematic disparities indicate potential bias that could translate into unequal clinical outcomes.

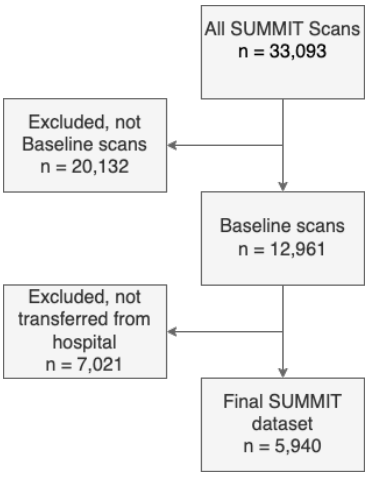

Figure 2: Flowchart of SUMMIT scan selection. Of 33,093 scans across all timepoints, 12,961 baseline scans were identified. At the time of the project, 5,940 had been transferred from hospital systems to the study research environment and were therefore available for analysis.

2. Methods

2.1 Study Design

Our study used a comparative design to examine whether there was a performance variation across sex and ethnicity in DL based nodule detection algorithms. Three state-of-the-art nodule detection algorithms were selected based on availability, diversity in architecture and performance on a classic lung nodule detection challenge, LUNA16. Initial training and testing was carried out using a subset of data drawn from the SUMMIT study, one of Europe's largest lung cancer screening studies. To further validate findings, a subset of data drawn from the Lung Screen Uptake Trial (LSUT) was used as an additional validation dataset.

2.2 Dataset Descriptions

SUMMIT. The SUMMIT study is a London based, lung cancer screening study that recruited high-risk participants. Participants were consented into the study between April 2019 and May 2021. Individuals aged 55 to 77 years with a history of smoking in north and east London were identified through electronic health records and invited through primary care. After initial eligibility was confirmed via a phone screener, qualifying participants were invited to attend a lung health check. During the lung health check, in-person eligibility assessments were conducted based on the 2013 United States Preventive Services Taskforce (USPSTF) Low-Dose CT screening criteria, and the 2012 Prostate, Lung, Colorectal, and Ovarian ($PLCO_{m2012}$) model. Eligible participants were offered LDCT. Participants were subsequently invited to attend annual lung health checks and interval scans over the following two years. Those without suspicious findings at year 1 were randomised to receive LDCT at year 1.

Sex was obtained from primary care records. Ethnic group was self-reported during the lung health check. Participants were selected from 16 categories, subdivided into five major groups, which aligns with the 2021 ONS Census: White (White British; White Irish; Other White); Asian (Bangladeshi; Indian; Pakistani; Other Asian); Black (Black African; Black Caribbean; Other Black); Mixed (White and Asian; White and Black African; White and Black Caribbean; Other Mixed); Other (Chinese; Any other ethnic group).

Written consent was obtained from participants at the point of determining study eligibility. Ethical approval was obtained from a NHS Research Ethics committee (17/LO/2004) and the NHS Health Research Authority's Confidentiality Advisory Group (18/CAG/0054).

A subset of 5,940 baseline LDCT scans was used in this analysis. This included all basline scans initially transferred from the hospital. Additional steps were taken to enrich the sample from underrepresented ethnic groups to support the intended analysis. This slightly enriched sample still reflects the broader SUMMIT cohort, which has a slight male predominance and a majority from the White ethnic group. The intersection of protected groups reveals a notable sex imbalance, particularly among Asian and Black females, whose numbers are too low for meaningful subgroup analysis. Figure 2 shows a flowchart outlining the selection process for SUMMIT scans.

All scans were captured using GE Revolution CT scanners by GE Healthcare, which were specially purchased for the SUMMIT study. The in-plane pixel spacing had a median of 0.674 mm^2 (IQR: 0.633-0.713 mm^2), ranging from

0.480 mm^2 to 0.977 mm^2. The slice thickness was highly consistent, with more than 99.9% of the scans with a slice thickness of 0.625 mm.

LSUT. The Lung Screen Uptake Trial (LSUT) is a London-based randomized controlled trial (RCT) of invitation methods to lung cancer screening, which began in April 2016. Participants were identified from primary care practices across London and invited to attend a pre-scheduled lung health check. The initial eligibility criteria for participants included being aged 60-75 and having a documented smoking history in primary care records within the past six years. Sex and ethnic group data were extracted from the primary care records, with ethnic group classifications following the 2011 Census guidelines. These ethnic categories were further categorized into 16 and subsequently 5 broad groups, as previously outlined for the SUMMIT dataset. Similar to SUMMIT, risk-based eligibility checks were carried out before, participants were deemed eligible for LDCT, these included $PLCO_{m2012}$ and Liverpool Lung Project (LLP v2.0). Of the 759 scans available in the validation dataset, only 297 were included due to necessary reprocessing under the SUMMIT protocol. This process led to the exclusion of cases that could not be processed by the CADe system. To ensure consistency, 146 scans originally labeled as containing nodules were reprocessed, while 151 scans initially reported as nodule-free were confirmed and retained. Cases with discrepancies between the original and reprocessed results were excluded due to resource constraints. Additionally, a small number of scans with an outlier-level nodule count were removed to maintain dataset integrity. Figure 3 illustrates the stepwise inclusion and exclusion process, beginning with all available LSUT scans and resulting in the final dataset used for analysis.

The scanners used for LSUT were from Canon Medical Systems, Aquilion series. The majority of scans were performed using the premium Aquilion ONE model. The mean pixel spacing across all of the scans is 0.677 mm^2 (IQR: 0.625-0.723 mm^2), ranging from 0.512 mm^2 to 0.976 mm^2. The majority of the scans have a slice thickness of 0.800 mm (89%), there are a small number of scans (11%) with a slice thickness of 0.300 mm.

Nodule annotations. For both datasets, nodule annotations were recorded using the same process. Each LDCT scan was reviewed by a specialist pulmonary radiologist, supported by Mevis'^{TM} Veolity CADe software. In addition to the location of nodules, maximum diameter and type were also recorded. It is important to note that pulmonary nodule annotation is subject to considerable inter- and intra-observer variability, as has been well documented in the literature Chen et al. (2022). While the use of a CADe system provides some guardrails and standardization to the annotation process (Singh et al. (2011)), variability in radiologist interpretation and nodule classification remains an inherent challenge in this domain.

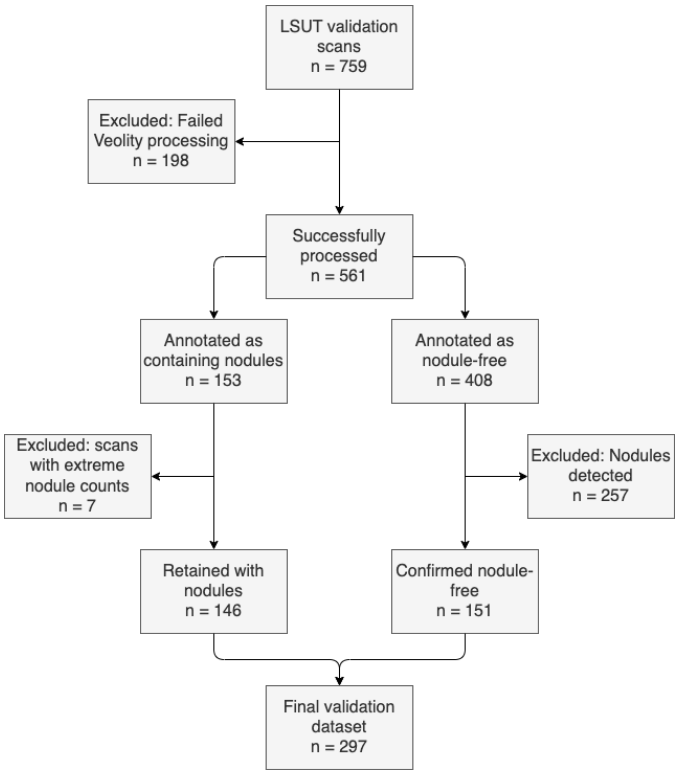

Figure 3: Flowchart of LSUT scan selection. Starting from 759 LSUT validation scans, exclusions were applied for outlier cases, failed Veolity processing, and inconsistencies between reprocessed and annotated findings. This process yielded a final dataset of 297 scans, comprising 147 with confirmed nodules and 151 confirmed nodule-free.

Participant overlap. The exclusion criteria for the SUMMIT study did not explicitly exclude participants from the LSUT study, so the potential remains that some LSUT participants could have also been part of SUMMIT. However, due to the anonymization of the data, there is no way to definitively identify any potential overlap between the two groups, however unlikely.

2.3 Nodule Detection Algorithms

Three open-source object detection frameworks were selected to assess the impact of model architecture on performance, particularly in the context of fairness in nodule detection. These frameworks employ different backbone networks, preprocessing pipelines, and loss functions, and each one has previously demonstrated state-of-the-art performance on the LUNA16 nodule detection challenge (Setio et al. (2017)). All models were implemented following the instructions provided by the authors. GitHub links to the code are included at the end of each description for reference.

Model 1: GRT123 U-Net. Model 1 was the winning entry in the Kaggle Data Science Bowl 2017 (Liao et al. (2019)). It is a one-stage object detector that employs a modified UNet (Ronneberger et al. (2015)) architecture as its backbone, with approximately 5 million trainable parameters. The preprocessing pipeline includes lung segmentation, resampling, and intensity clipping within the range of -1200 to 300 Hounsfield Units (HU). The loss functions used for this model were binary cross-entropy for nodule classification and smoothed L1 loss for bounding box regression. To address the underrepresentation of larger nodules, the original training strategy included targeted oversampling, where nodules with diameters between 30 mm and 40 mm were duplicated to generate twice as many training samples, while nodules measuring 40 mm or greater were replicated fourfold. This ensured that each of these nodule categories contributed proportionally more examples during training.[1]

Model 2: MONAI RetinaNet. Model 2 is the MONAI detection algorithm, which employs a RetinaNet (Lin et al. (2018b)) architecture with approximately 21 million trainable parameters. It incorporates a feature pyramid network (Lin et al. (2017)) to facilitate multi-scale detection and utilizes Focal Loss (Lin et al. (2018a)) to address class imbalance. The preprocessing pipeline includes resampling and intensity clipping within the range of -1024 to 300 Hounsfield Units (HU).[2].

Model 3: TiCNet Transformer. Model 3 TiCNet (Ma et al. (2024)), comprises three main parts, a 3D Encoder-Decoder network, a Transformer Encoder-Decoder module and a two-head nodule decoder. Key innovations include incorporating attention blocks into the 3D Encoder-Decoder to focus on the most important feature channels and the Transformer module to help in capturing long-range dependencies (all-be-it, within the context of patch-based training). The two-head nodule decoder integrates a false positive reduction arm has not be used as part of this project, this is to ensure fair comparison with the other two models. The final model had approximately 28 million trainable parameters.[3]

All three models use hard-negative mining (HNM) (Bucher et al. (2016)) and all three models use a patch-based training method (128^3 for GRT123 U-Net and 3, $198 \times 198 \times 80$ for MONAI RetinaNet).

The output from all three models consists of a list of candidate locations for each scan. Each candidate is defined by its centroid and diameter, corresponding to the potential 'nodule' at that location, along with a probability indicating the likelihood that it is a nodule.

1. https://github.com/lfz/DSB2017/tree/master)
2. https://github.com/Project-MONAI/tutorials/tree/main/detection
3. https://github.com/MIAinCS/TiCNet/

Model Comparison Summary. For ease of comparison, we present Table 1, which displays the key characteristics of all three models side by side. This table summarizes the preprocessing steps, network architectures, and loss functions for each model, allowing a straightforward comparison of their main features. All models were implemented as released, without adaptation.

2.4 Evaluation Metrics

To ensure a clinically meaningful and equitable evaluation of model performance, we adopt a two-level evaluation strategy: (1) detection-oriented nodule-level metrics, which traditionally benchmark the model's ability to detect and classify actionable nodules based on established criteria, and (2) fairness-oriented scan-level metrics, which focus on the detection of actionable nodules across demographic groups to assess equality of clinical outcomes. At both levels, we evaluate performance specifically for actionable nodules, as these are the nodules that would trigger follow-up actions in a clinical setting. The nodule-level approach is the traditional method for evaluating detection accuracy, while the scan-level metrics provide an additional layer aimed at ensuring equitable clinical outcomes for all demographic groups.

Nodule-Level Detection Metrics. The primary performance metric aligns with that defined in the LUNA16 challenge as the Competition Performance Metric (CPM). It relies on the Free-Response Receiver Operating Characteristic (FROC) curves that measure sensitivity at different levels of false positive per scan. The CPM is calculated as the average across 7 fixed operating points of false positives per scan, namely ($\frac{1}{8}$, $\frac{1}{4}$, $\frac{1}{2}$, 1, 2, 4, 8). A nodule is considered positively detected if the centroid of the predicted nodule lies within the radius of the annotated nodule centroid.

Since the FROC analysis does not account for the predicted nodule diameter, an important factor in determining whether a nodule is actionable, we complement it by also evaluating mean Average Precision (mAP). The mAP incorporates the predicted nodule diameter by using Intersection over Union (IoU) thresholds of 0.1, 0.3, and 0.5, providing a more complete assessment of model performance.

For both FROC/CPM and mAP, 95% confidence intervals were computed using bootstrapping with 1000 bootstraps. For each bootstrap, a new set of candidate locations are generated by sampling scans with replacement.

As bootstrapping with replacement is used to generate the FROC and CPM metrics, statistical significance will be evaluated using a 95% confidence interval for the difference in CPM between groups. If the confidence interval crosses zero, this indicates that the difference is not statistically significant.

Table 1: Summary of preprocessing, architectures, and training loss functions for each model. All implementations were used out-of-the-box without adaptation, following their original published or released configurations.

Aspect	GRT123 U-Net	MONAI RetinaNet	TiCNet Transformer
Patch Size	128^3	$198 \times 198 \times 80$	96^3
Parenchymal segmentation	Yes	No	Yes
Architecture	3D U-Net	3D RetinaNet (FPN)	CNN + Transformer
Cls. Loss	BCE	Focal	Weighted BCE
Reg. Loss	Smooth L1	Smooth L1	Smooth L1
Other	HNM, oversample	HNM	Attention, HNM

Scan-Level Metrics. Fairness is assessed at the scan level, where a scan is considered positive if it contains at least one actionable nodule. A positive prediction for such a scan is counted as a true positive if any actionable nodule is detected. Likewise, a scan with no actionable nodules is considered a true negative if no predictions are made for that scan.

We report the following metrics across protected groups to assess fairness:

- **True Positive Rate (TPR)**: The proportion of positive scans that have at least one correctly detected actionable nodule. Similar TPRs across groups reflect equal opportunity.
- **False Positive Rate (FPR)**: The proportion of negative scans that have at least one false positive prediction. Similar FPRs across groups reflect equal burden of erroneous call.
- **Positive Predictive Value (PPV)**: The proportion of scans predicted to be positive that are truly positive.
- **Prediction Rate (PR)**: The proportion of all scans that receive at least one positive prediction.

To calculate these metrics, we evaluate model predictions across a range of decision thresholds. We mirror the nodule-level analysis, by computing each metric at the 7 fixed false positive rates per scan, then averaging the result to produce a single summary value. Confidence intervals [95%] were estimated using 1,000 bootstrap resamples with replacement at the scan level and are reported as error bars. As with the nodule-level analysis, statistical significance was assessed separately by bootstrapping the difference in means between groups and examining whether the resulting 95% confidence interval crossed zero.

To help interpret the results, we also report the number of nodules per scan in each demographic group, since scans with more nodules may be easier to classify correctly and can influence metric values. Standardised residuals of the associated chi-square hypothesis testing are derived as $(\text{observed count} - \text{expected count})/\sqrt{\text{expected count}}$.

2.5 Model Calibration

To check whether model calibration differs across demographic groups, we compute the Expected Calibration Error (ECE) separately for each category within the sex and ethnic group attributes. This allows us to determine whether the model's confidence scores are well-calibrated for all categories or if there are systematic differences between them.

Since this is an object detection task rather than a classification task, we use Precision as a proxy for Accuracy, following the approach detailed in Kuppers et al. (2020). ECE measures the alignment between predicted confidence scores and actual detection performance. If the calibration error varies across categories within each protected group, it may indicate that the model's confidence estimates are not equally reliable across different demographic groups.

2.6 Experiments

Our experiments were designed to systematically assess how various types of imbalance in the training data influence fairness in algorithm performance. Experiment 1 establishes a baseline by training on all available data, thereby replicating the sex and ethnicity distributions typical of real-world screening datasets. Experiment 2 removes sex as a potential confounder, isolating the impact of ethnic diversity on performance. Experiment 3 does the reverse, focusing on a single ethnic group to examine the role of sex balance in training. Finally, Experiment 4 explores key drivers of model performance, providing insight into which factors most strongly influence disparities across demographic groups.

Experiment 1: Balanced Ethnicity Test Set. The goal of the first experiment, which utilized all available data, was to maximize the training dataset while preserving the original, imbalanced distribution of sex and ethnic groups, thereby reflecting real-world screening data. To enable meaningful comparisons, we constructed a test set that ensured a minimum number of cases per ethnic group, balancing Asian, Black, and White participants so that con-

fidence intervals would be sufficiently narrow for statistical evaluation.[4]

Experiment 2: Male-Only Dataset. The second experiment, which with the third experiment aimed to mitigate specific confounding factors, isolated the impact of ethnic diversity on algorithmic performance while eliminating sex as a confounder. Given the under representation of females in certain ethnic groups, we restricted the training and testing dataset to male patients, ensuring a test set balanced for ethnic groups. As in the first experiment, the training dataset had a similar sample distribution to the total SUMMIT dataset.[5]

Experiment 3: White-Only Dataset. Experiment 3 assessed the impact of sex representation in training while reducing ethnic group, confounding factors, by training exclusively on a single ethnicity. Sampling objectives remained consistent with the previous experiments, aligning with underlying sample proportions while ensuring sufficient, balanced test data. The White ethnic group was selected to provide an adequate amount of training data.

We validate the findings by comparing the relative performance of each group in the validation dataset (LSUT), using models trained in experiment 1(Test Balanced). Performance is visualized across sex and ethnic groups. Due to the low number of non-white ethnic groups in the LSUT dataset, ethnic categories were collapsed into White and Other.

Experiment 4. Finally, in the fourth experiment, post-hoc analysis was carried out, designed to explore factors influencing nodule detection algorithm performance. Specifically, it examines the impact of nodule characteristics such as nodule type and diameter on performance. Nodule diameter was categorized into five ranges: 0-6mm, 6-8mm, 8-30mm, 30-40mm, and 40+mm, aligning with the SUMMIT nodule management protocol, which in turn is based on the British Thoracic Society (BTS) guidelines for the management of pulmonary nodules (British Thoracic Society (2015)), with additional categories (30-40mm and 40+mm) included to address special cases, such as oversampling in GRT123 U-Net. These findings are validated on the LSUT dataset. A key focus of this analysis is understanding the training set composition, particularly the prevalence of each nodule type and the distribution of nodule sizes, and how these factors influence model performance.

4. The available counts were: Asian = 666, Black = 454, White = 4485, Mixed = 128, Other = 206. Mixed and Other were retained in training but excluded from testing due to insufficient sample sizes for reliable subgroup analysis.
5. Female-only subgroup analyses were not feasible because the available counts by ethnicity were highly imbalanced (Asian = 57, Black = 172, Mixed = 59, Other = 57, White = 1991). Preserving training sets that reflected the overall SUMMIT distribution while also constructing balanced test sets would have left only very small numbers of Asian, Mixed, and Other females, making meaningful comparisons across groups unreliable. Restricting this experiment to males ensured sufficient representation of each ethnic group for consistent evaluation.

3. Results

The initial sections of the results compare model performance across sex and ethnic groups at both the nodule and scan level. For the nodule-level analysis, we report sensitivities with 95% confidence intervals at fixed false-positive rates, alongside mean Average Precision (mAP) at multiple IoU thresholds. At the scan level, we summarize performance using error-bar plots of key metrics (TPR, FPR, PPV, PR) with bootstrapped confidence intervals. To assess disparities, we include plots of group-wise differences with 95% confidence intervals. Table 2 provides the demographic composition of the datasets, and Table 3 summarizes the distribution of nodules across groups, highlighting an underrepresentation of Black individuals in the '2+ nodules' category. These descriptive patterns provide important context for interpreting performance differences.

3.1 Experiment 1: Balanced Ethnicity Test Set

Table 4 summarizes the demographic composition of the balanced ethnicity test set.

Nodule-Level Anlaysis. Nodule-level results are shown in Figure 4. Across models, male and female CPM scores were very similar (see Supporting Table S1a), with all confidence intervals overlapping zero (Figure 5). The largest gap was observed for GRT123 U-Net, where CPM was 0.41 [95% CI: 0.32-0.51] for males and 0.46 [95% CI: 0.36-0.57] for females, but this difference was not statistically significant.

In contrast, ethnic differences were more pronounced: White participants achieved higher CPM, e.g. 0.49 [95% CI: 0.39-0.60] vs. 0.35 [95% CI: 0.23-0.48] for Asian participants on the same model, with non-overlapping CIs indicating significant disparity. Black participants consistently fell between White and Asian, typically around 0.44-0.45.

The mAP scores (Figure 6) showed slightly different results between ethnic groups, White participants still demonstrated the best performance across all models, but Asian and Black participants were much more closely aligned, with only slightly overlapping confidence intervals compared to White participants.

Scan-Level Analysis. At the scan level (Figure 7), sex differences were negligible. For example, TPR was 0.54 [95% CI: 0.44-0.63] for females vs. 0.52 [95% CI: 0.42-0.60] for males in GRT123 U-Net, and similar overlaps were observed across models. The only consistent sex-based trend was a slightly higher PPV for females, e.g. 0.30 vs.

Table 2: Sex and Ethnicity proportions in SUMMIT

Group	Category	SUMMIT Subset	Total SUMMIT	LSUT Subset	Total LSUT
Sex	Female	2336 (39.3%)	5508 (42.5%)	123 (41.4%)	336 (44.1%)
	Male	3604 (60.7%)	7450 (57.5%)	174 (58.6%)	426 (55.9%)
Ethnic	Asian	666 (11.2%)	845 (6.5%)	1 (0.3%)	7 (0.9%)
Group	Black	454 (7.6%)	577 (4.5%)	26 (8.8%)	77 (10.1%)
	Mixed	128 (2.2%)	283 (2.2%)	5 (1.7%)	8 (1.0%)
	Other	206 (3.5%)	451 (3.5%)	12 (4.0%)	31 (4.1%)
	White	4486 (75.5%)	10802 (83.4%)	253 (85.2%)	636 (83.5%)
Total		5940 (100%)	12958 (100%)	297 (100%)	759 (100%)

Table 3: Distribution of nodule counts and standardized residuals, across sex and ethnic groups in the SUMMIT cohort.

Group	Category	0 Nodules	1 Nodule	2+ Nodules
Sex	Female	4230 (−0.26)	840 (−0.31)	440 (1.27)
	Male	5760 (0.22)	1157 (0.26)	534 (−1.1)
Ethnic	Asian	647 (−0.2)	137 (0.58)	62 (−0.2)
Group	Black	466 (1.01)	82 (−0.73)	**29 (−2.18)**
	Mixed	221 (0.19)	42 (−0.24)	20 (−0.27)
	Other	365 (0.93)	59 (−1.26)	27 (−1.18)
	White	8291 (−0.4)	1677 (0.3)	836 (0.85)

Table 4: Profile of protected groups - Experiment 1: Balanced Ethnicity Test Set

Group	Category	Training	Validation	Test
Sex	Female	1961 (38.6%)	125 (46.8%)	250 (42.1%)
	Male	3118 (61.4%)	142 (53.2%)	344 (57.9%)
Ethnic	Asian	443 (8.7%)	25 (9.4%)	198 (33.3%)
Group	Black	244 (4.8%)	12 (4.5%)	198 (33.3%)
	Mixed	119 (2.3%)	9 (3.4%)	0 (0%)
	Other	199 (3.9%)	7 (2.6%)	0 (0%)
	White	4074 (80.2%)	214 (80.1%)	198 (33.3%)
Total		5079 (100.0%)	267 (100.0%)	594 (100.0%)

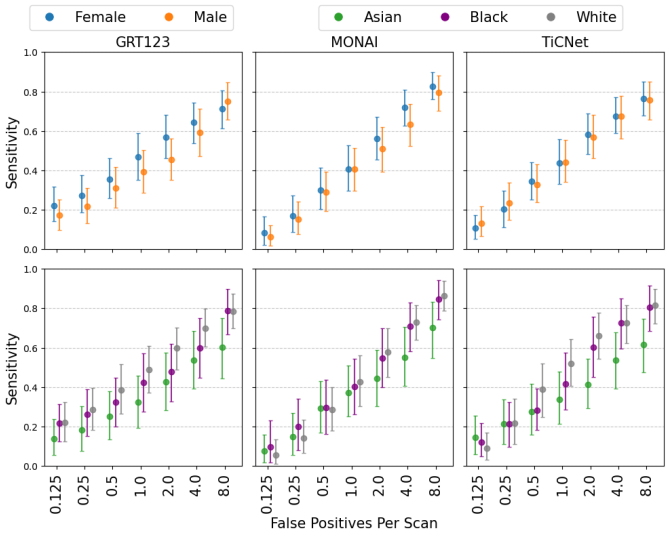

Figure 4: Experiment 1 (Balanced Ethnicity Test Set), error bar plots. Sensitivity for 7 fixed false positive per scan rates, with error bars generated from 1000 bootstraps for each model across protected groups. Top: comparison across sex categories, Bottom: across ethnic group categories.

0.23 in GRT123 U-Net, though not statistically significant.

Ethnic group differences were clearer. White participants achieved higher TPR and PPV, for instance 0.63 [95% CI: 0.53-0.72] and 0.32 [95% CI: 0.24-0.41] vs. 0.42 [95% CI: 0.31-0.53] and 0.23 [95% CI: 0.15-0.32] for Asian participants. Black participants again showed intermediate performance, e.g. TPR of 0.49 [95% CI: 0.36-0.62].

Confidence interval analysis confirmed significant disparities for White-Asian comparisons, while Black-White and Black-Asian contrasts were generally non-significant (see Supporting Figure S1). Nodule count distributions by demographic group are given in Supporting Table S1b and show that White participants have a significantly higher number of scans with $2+$ nodules.

3.2 Experiment 2: Male-Only

The sample composition for Experiment 2 (Male-Only) is summarized in Table 5.

Nodule-level analysis. As shown in Figure 8, Asian participants achieved the highest CPM scores (see Supporting Table S2a), for two of the three models (e.g. 0.46 [95% CI: 0.31-0.61] for GRT123 U-Net), despite being underrepresented in training, while White individuals showed similar CPM values (e.g. 0.43 [95% CI: 0.31-0.57]). Black individuals had consistently lower CPMs (e.g. 0.33 [95% CI: 0.19-0.50]).

Confidence interval overlap (see Supporting Figure S2a) indicates no statistically significant differences.

Mean Average Precision (mAP) scores (see Supporting Figure S2b), however, consistently favored White cases (e.g.

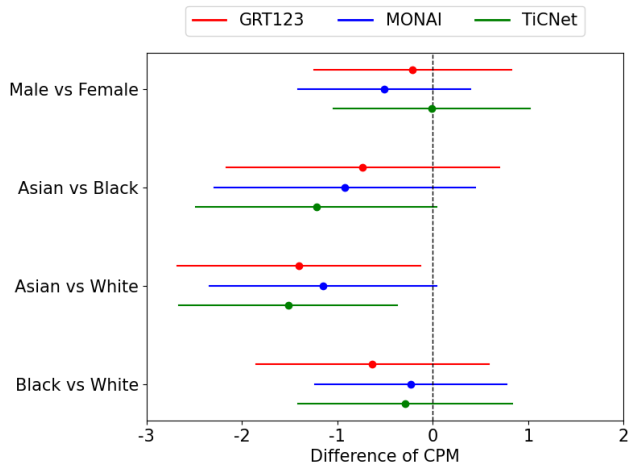

Figure 5: Experiment 1 (Balanced Ethnicity Test Set), Confidence Interval for Difference in CPM (1000 bootstraps). Sex and ethnic group, category comparisons, across all models.

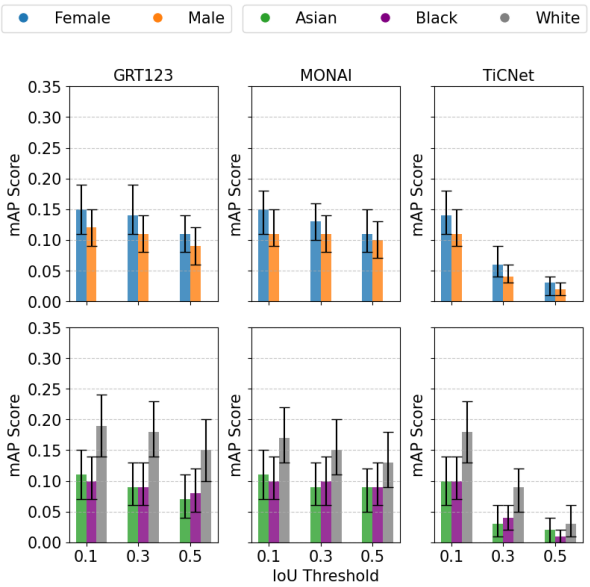

Figure 6: Experiment 1 (Balanced Ethnicity Test Set), mAP scores for IoU thresholds of 0.1, 0.3, and 0.5. Error bars generated from 1000 bootstraps. Top: comparison across sex categories, Bottom: across ethnic groups.

0.20 [95% CI: 0.14-0.26] in MONAI RetinaNet) over Asian and Black participants (both ≈ 0.10).

Scan-level analysis. The scan-level metrics show consistent patterns across all models (Figure 9). While White participants continue to demonstrate the strongest performance on most metrics, Asian participants now achieve comparable results, particularly in TPR and PR. For example, with MONAI RetinaNet, TPR for White and Asian groups are 0.61 [95% CI: 0.51-0.70] and 0.54 [95% CI: 0.41-0.68], respectively, and PR values are similarly close. In contrast, Black participants exhibit a noticeable drop

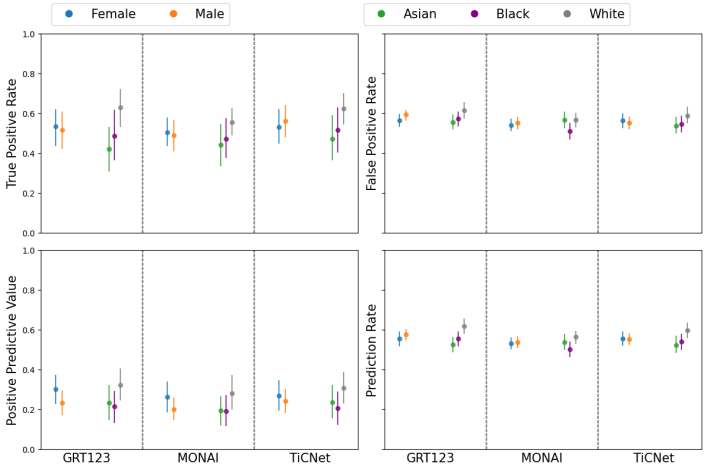

Figure 7: Scan-level performance metrics by sex and ethnic group for Experiment 1 (Balanced Ethnicity Test Set). Each subplot shows the mean and 95% confidence interval for TPR (top left), FPR (top right), PPV (bottom left), and PR (bottom right), averaged across seven fixed false-positive-per-scan thresholds. Error bars represent 95% confidence intervals from 1000 bootstraps.

Table 5: Profile of protected groups - Experiment 2 (Male-Only)

Ethnic Group	Training	Validation	Test
Asian	98 (6.2%)	11 (10.5%)	140 (33.3%)
Black	71 (4.5%)	4 (3.8%)	140 (33.3%)
Mixed	34 (2.2%)	2 (1.9%)	0 (0%)
White	1317 (83.7%)	84 (80.0%)	140 (33.3%)
Other	53 (3.4%)	4 (3.8%)	0 (0%)
Total	1573 (100.0%)	105 (100.0%)	420 (100.0%)

in performance across all metrics, with MONAI RetinaNet TPR at 0.46 [95% CI: 0.37-0.55] and PPV at 0.15 [95% CI: 0.09-0.24]. These trends are consistent across all evaluated models. Notably, the differences in scan-level metrics between Black participants and other groups are statistically significant for several metrics, as shown in Supporting Figure S2c.

Residual analysis (Table 6) suggested a trend toward more multi-nodule scans in White participants (+1.94) compared to Asian (-0.51) and Black (-1.43).

Table 6: Distribution of nodule counts per scan and standardized residuals by ethnic group in Experiment 2 (Male-Only) test set.

Ethnic Group	0 Nodules	1 Nodule	2+ Nodules
Asian	110 (0.39)	21 (-0.48)	9 (-0.51)
Black	113 (0.68)	21 (-0.48)	6 (-1.43)
White	95 (-1.07)	28 (0.97)	17 (1.94)

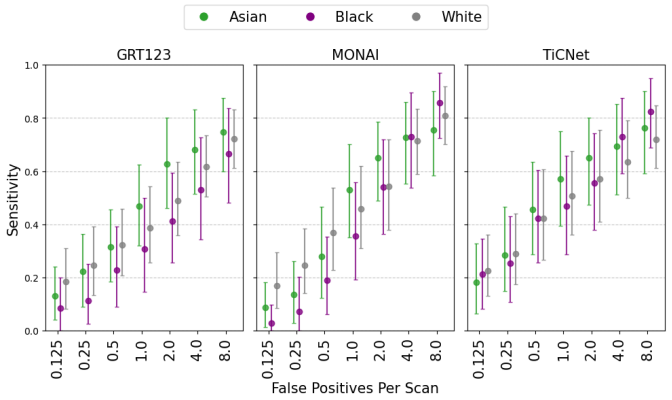

Figure 8: Experiment 2 (Male-Only), error bar plots. Sensitivity for 7 fixed false positive per scan rates, with error bars generated from 1000 bootstraps for each model across ethnic groups.

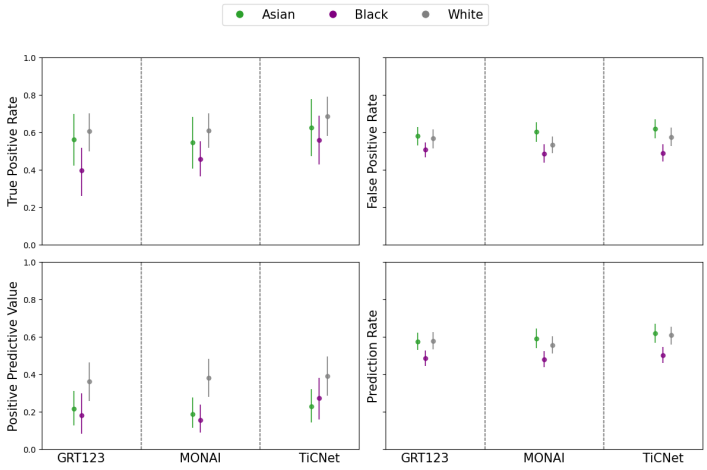

Figure 9: Scan-level performance metrics by ethnic group for Experiment 2 - (Male-Only). Each subplot shows the mean and 95% confidence interval for TPR (top left), FPR (top right), PPV (bottom left), and PR (bottom right), averaged across seven fixed false-positive-per-scan thresholds. Error bars represent 95% confidence intervals from 1000 bootstraps.

3.3 Experiment 3: White-Only

The breakdown of the sample used for Experiment 3 (White-Only) is shown in Table 7.

Table 7: Profile of protected groups - Experiment 3 (White-Only)

Sex	Training	Validation	Test
Female	1494 (44.4%)	98 (43.8%)	399 (50.0%)
Male	1870 (55.6%)	126 (56.2%)	399 (50.0%)
Total	3364 (100.0%)	224 (100.0%)	798 (100.0%)

Nodule-level analysis. The results from training on a white participants, only sample reveal small and inconsistent differences in performance between male and female groups across the models (Figure 10). For example, the CPM (see Supporting Table S3a) for males was marginally higher than for females for GRT123 U-Net, but the opposite pattern was seen for MONAI RetinaNet and TiCNet Transformer. None of the differences were statistically significant, 95% CI overlapping zero for all models (see Supporting Figure S3a).

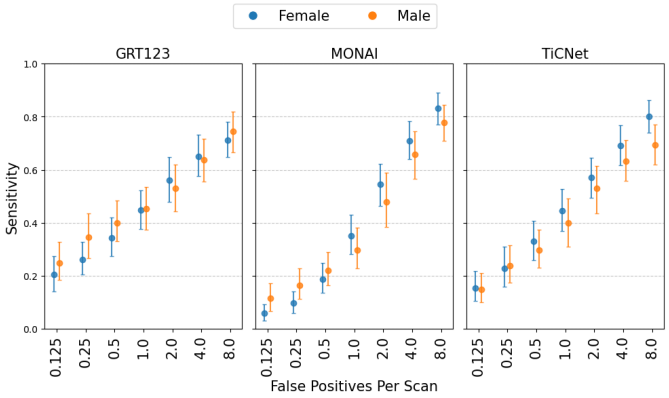

Figure 10: Experiment 3 (White-Only), error bar plots. Sensitivity for 7 fixed false positive per scan rates, with error bars generated from 1000 bootstraps for each model across sex.

In contrast, the mAP scores (Figure 11) showed slightly better performance for females across all models and IoU thresholds, but again confidence intervals overlapped, indicating no significant differences.

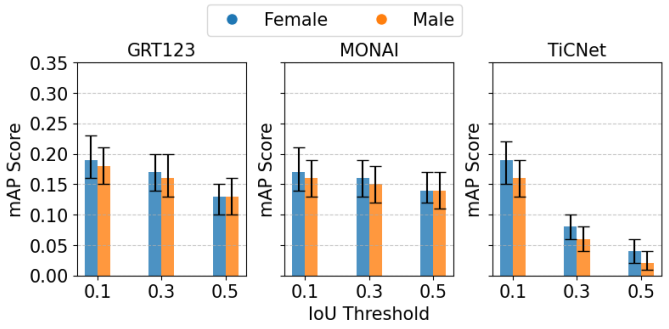

Figure 11: Experiment 3 (White-Only), mAP scores for IoU thresholds of 0.1, 0.3, and 0.5, with error bars generated from 1000 bootstraps.

Scan-level analysis. Scan-level metrics (Figure 12) showed fluctuations across models but no consistent advantage for either sex. For instance, for MONAI RetinaNet, FPR was 0.50 [95% CI: 0.47-0.53] for females vs. 0.52 [95% CI: 0.49-0.55] for males, and PPV was 0.32 [95% CI: 0.25-0.39] vs. 0.30 [95% CI: 0.24-0.37]. Other models showed similar overlapping confidence intervals (see Supporting Figure S3b). Only GRT123 U-Net indicated a possible male advantage

in FPR and PR.

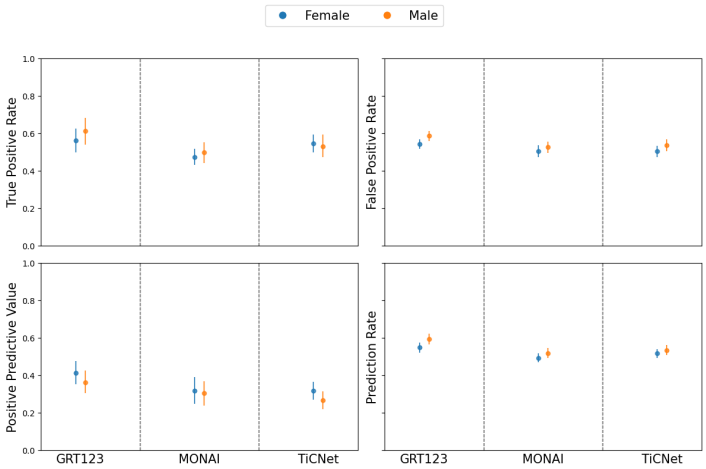

Figure 12: Scan-level performance metrics TPR (top-left), FPR (top-right), PPV (bottom-left) and PR (bottom-right) across sex for Experiment 3 (White-Only).

Table 8 presents the distribution of the number of nodules per scan across male and female participants in Experiment 3. The table shows no significant deviations from the expected distribution.

Table 8: Distribution of nodule counts per scan and standardized residuals by sex in Experiment 3 (White-Only) test set.

Sex	0 Nodules	1 Nodule	2+ Nodules
Female	259 (-0.52)	89 (0.27)	51 (0.89)
Male	276 (0.52)	84 (-0.27)	39 (-0.89)

3.4 External Validation

Inference on the LSUT dataset was conducted using models trained on data from Experiment 1 (Balanced Ethnicity Test Set). This approach was consistent across all three models, GRT123 U-Net, MONAI RetinaNet, and TiCNet Transformer.

Nodule-level analysis. The overall CPM performance (see Supporting Table S4a) for each of the three models was compared across the SUMMIT and LSUT datasets, with results shown in Figure 13. Both GRT123 U-Net and MONAI RetinaNet exhibited moderate improvements on LSUT: GRT123 U-Net increased slightly from 0.44 [95% CI: 0.37-0.51] to 0.47 [95% CI: 0.39-0.56], while TiCNet Transformer improved from 0.44 [95% CI: 0.38-0.51] on SUMMIT to 0.53 [95% CI: 0.44-0.62] on LSUT, a gain of 0.09. Notably, MONAI RetinaNet demonstrated the largest improvement, rising from 0.42 [95% CI: 0.36-0.49] on SUMMIT to 0.55 [95% CI: 0.46-0.63] on LSUT, a substantial gain of 0.13. These results suggest that all models generalised well to external validation, with MONAI RetinaNet showing the strongest performance on the LSUT dataset.

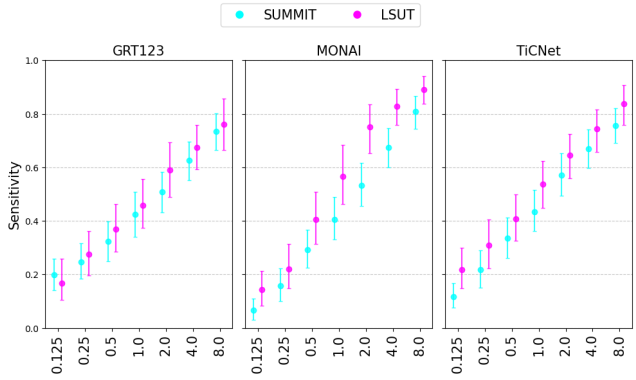

Figure 13: Comparison of model performance between the primary dataset (SUMMIT) and the external validation dataset (LSUT). Sensitivity for 7 fixed false positive per scan rates, with error bars generated from 1000 bootstraps. Inference was conducted using the model trained in Experiment 1 (Balanced Ethnicity Test Set).

Group-specific comparisons (Figure 14) showed patterns broadly consistent with the SUMMIT analyses between males and females where the differences in CPM values were negligible, with confidence intervals centered around zero (see Supporting Figure S4a).

For ethnicity, the LSUT dataset required collapsing Asian, Black, Mixed, and Other into one group due to low sample sizes. Interestingly, this group outperformed White participants across all models (CPM 0.56-0.72 vs. 0.46-0.52). This contrasts with SUMMIT, where White participants generally had higher CPM. Confidence intervals indicated a statistically significant CPM difference between White and Other groups in LSUT.

For the mAP scores (see Supporting Figure S4b), males slightly outperform females, but confidence intervals overlap, indicating no significant difference. For ethnicity, GRT123 U-Net shows higher mAP for White cases, while MONAI RetinaNet and TiCNet Transformer favour the group comprising of Asian, Black, Mixed and Other ethnic groups, consistent with CPM results; however, all differences lack statistical significance due to overlapping confidence intervals.

Scan-level analysis. Scan-level metrics (Figures 15-16) generally showed improved performance on LSUT compared to SUMMIT. For example, TiCNet Transformer achieved higher TPR (0.75 vs. 0.55) and PPV (0.48 vs. 0.25). Differences in PR and FPR were smaller (e.g., GRT123 U-Net PR 0.60 vs. 0.57).

Across sex groups, performance was consistent (e.g., TiCNet TPR 0.77 females vs. 0.73 males). Ethnic group

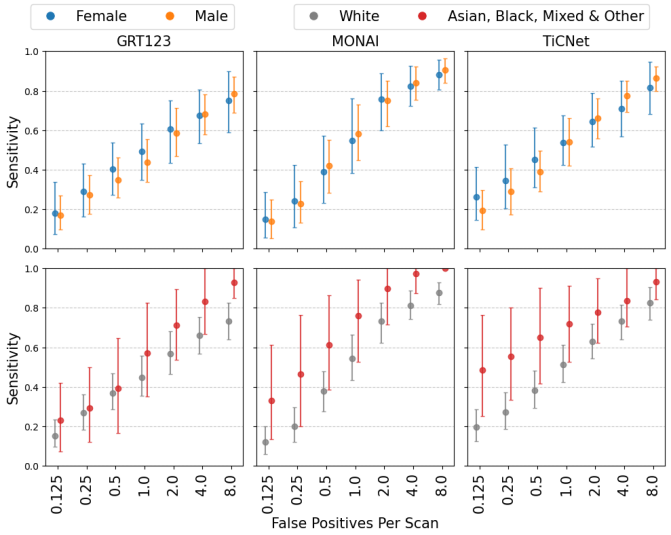

Figure 14: Error plots, generated from running inference on external dataset (LSUT), using models trained in Experiment 1 (Balanced Ethnicity Test Set). Sensitivity for 7 fixed false positive per scan rates, with error bars generated from 1000 bootstraps for each model across protected groups on the LSUT dataset. The top plot compares sex groups, while the bottom plot compares ethnicities.

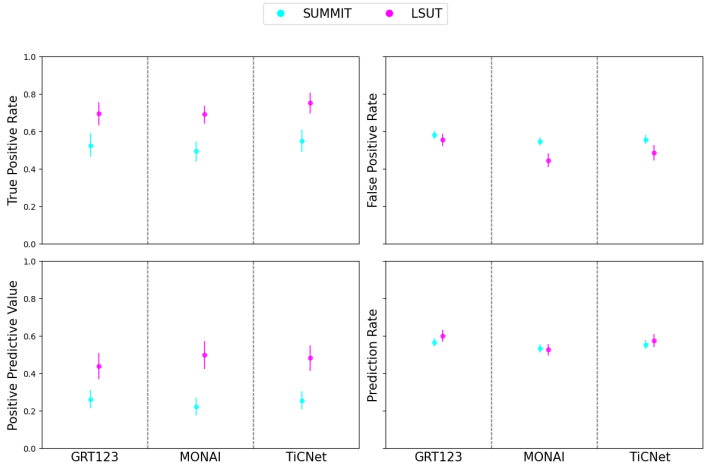

Figure 15: Comparison of scan-level metrics across models and datasets. Each panel corresponds to one of the three models evaluated. TPR, FPR, PPV, and PR are averaged across seven fixed false positive per scan thresholds.

comparisons showed greater variability, with participants in the combined Asian, Black, Mixed and Other ethnic group achieving slightly higher TPR and PPV, but confidence intervals overlapped, suggesting no systematic bias (see Supporting Figure S4c).

The distribution of nodule counts by sex and ethnicity in the LSUT subsample (see Supporting Table S4b) shows standardized residuals within ± 2, indicating no significant deviations from expected values. This suggests balanced representation of nodule categories across groups.

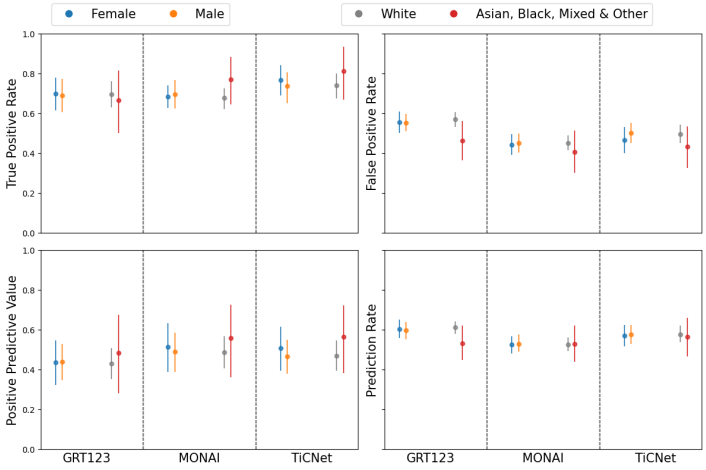

Figure 16: Scan-level metrics, generated from running inference on external dataset (LSUT), using models trained in Experiment 1 (Balanced Ethnicity Test Set). Each metric is calculated as an average across 7 false positive per scan rates and the error bars are generated from 1000 bootstraps. Sex and ethnic group plots are shown per Model

3.5 Model Calibration

The calibration for each model for the data used in Experiment 1 (Balanced Ethnicity Test Set) is presented in Figure 17. The results show that all models are poorly calibrated, with precision only increasing sharply at high confidence levels, suggesting that the models tend to be overconfident in their predictions. However, MONAI RetinaNet demonstrates significantly better calibration than GRT123 U-Net or TiCNet Transformer, with an overall ECE of 0.349, compared to 0.580 for GRT123 U-Net and 0.713 for TiCNet Transformer. Calibration patterns were consistent across sex and ethnic group.

3.6 Post-Hoc analysis: Drivers of Algorithm Performance

This analysis evaluates the sensitivity of each model based on nodule characteristics, specifically nodule type and nodule diameter. As shown in Figure 18, sensitivity is plotted against the number of false positives per scan for consistency. Additionally, we provide the distribution of nodules with each characteristic observed during training for reference.

MONAI RetinaNet consistently outperforms GRT123 U-Net and TiCNet Transformer across the three dominant size categories (0-6mm, 6-8mm, and 8-30mm), which also make up the majority of the test set. Notably, MONAI RetinaNet exhibits a particularly strong advantage in detecting the smallest nodules (0-6mm), which are also the most prevalent, contributing significantly to its superior overall performance. For larger nodules (30-40mm and 40mm+), MONAI RetinaNet shows slightly reduced detection rates at the lowest operating points compared to GRT123 U-Net and TiCNet Transformer, although this has a relatively

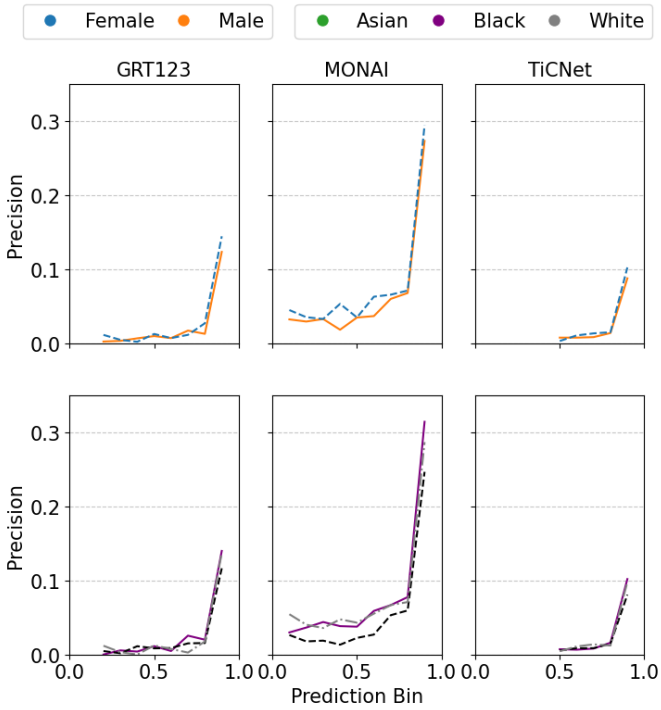

Figure 17: Precision-based calibration plots for three nodule detection models across different confidence bins. The top row presents calibration curves stratified by sex while the bottom row shows calibration by ethnicity (Black, Asian, White) in the Experiment 1 (Balanced Ethnicity Test Set) test set.

minor impact on overall performance due to their lower prevalence.

In the nodule type analysis, all three models show the highest performance for solid nodules, which are the most prevalent in both the training and test datasets, and therefore drive overall performance. The detection trends for other nodule types largely reflect their representation in the training data. However, part-solid nodules, despite being the least represented type in both datasets, show the second-best detection performance.

GRT123 U-Net differs from Models 2 and 3 in that it demonstrates higher performance at the lowest operating points for all nodule types and diameter categories. However, the overall performance of GRT123 U-Net is affected by the fact that, unlike Models 2 and 3, its performance increases more gradually at higher operating points. As a result, despite its initial strong detection rates, the shallower increases across operating points ultimately lead to a lower overall performance compared to the other models, which show more consistent and steeper improvements as the operating points increase.

LSUT. Figure 19 illustrates the relative performance of nodule characteristics for the LSUT dataset. The results mirror those observed with the SUMMIT dataset, with MONAI RetinaNet demonstrating better overall performance, excelling in the detection of smaller nodules. Specifically, MONAI RetinaNet shows a stronger performance for the 0-6 mm nodule category, with better performance at the lower operating points. While GRT123 U-Net and TiCNet Transformer perform better at detecting larger nodules, such as those in the 40+ mm category, particularly at low false positives per scan, these will have minimal impact on overall performance due to low numbers in the test dataset.

As with SUMMIT, for all three models, the more common nodule types seen during training exhibit steady and consistent performance, with detection rates increasing as the false positive per scan rate increases. Notably, part-solid nodules outperform non-solid nodules, even though part-solid types were less represented in both the training and test datasets. It is important to note that no actionable potentially infective nodules were recorded in the LSUT dataset.

Figures 20 and 21 further compare the composition of the test datasets across the SUMMIT and LSUT datasets, displaying the percentage distribution of nodules by type and diameter, respectively. A Chi-square test confirms a statistically significant difference in nodule type composition between the datasets ($\chi^2 = 110.91$, $P < 1e{-}19$). Additionally, nodule diameter distribution also differs significantly between the two datasets ($\chi^2 = 30.19$, $P = 4.47e-06$).

4. Discussion

In this study, we assessed variation in performance across sex and ethnic groups using three state-of-the-art pulmonary nodule detection algorithms trained on an unbalanced dataset from a large lung cancer screening programme. Scan-level metrics showed some variation in performance across the different experiments. White participants, the majority group within the training data, demonstrated significantly better performance than Asian participants in Experiment 1 (Balanced Ethnicity Test Set). However, when the confounding factor of sex was removed in Experiment 2 (Male-Only), performance became much more balanced, with no systemic differences observed across groups within this male-only cohort. It is important to note that we were not able to extend this analysis to a female-only subgroup, or to fully factorial combinations of sex and ethnicity, due to the small sample sizes of minority groups once stratified. Likewise, Experiment 3 (White-Only) isolates ethnicity effects only within white participants, and results cannot be assumed to generalise to other ethnic groups. A more complete intersectional analysis remains an important direction for future work as larger datasets become available.

Rather than performance being tied to a specific demographic group, our analysis suggests that scan-level metrics are, in part, influenced by the distribution of nodule counts

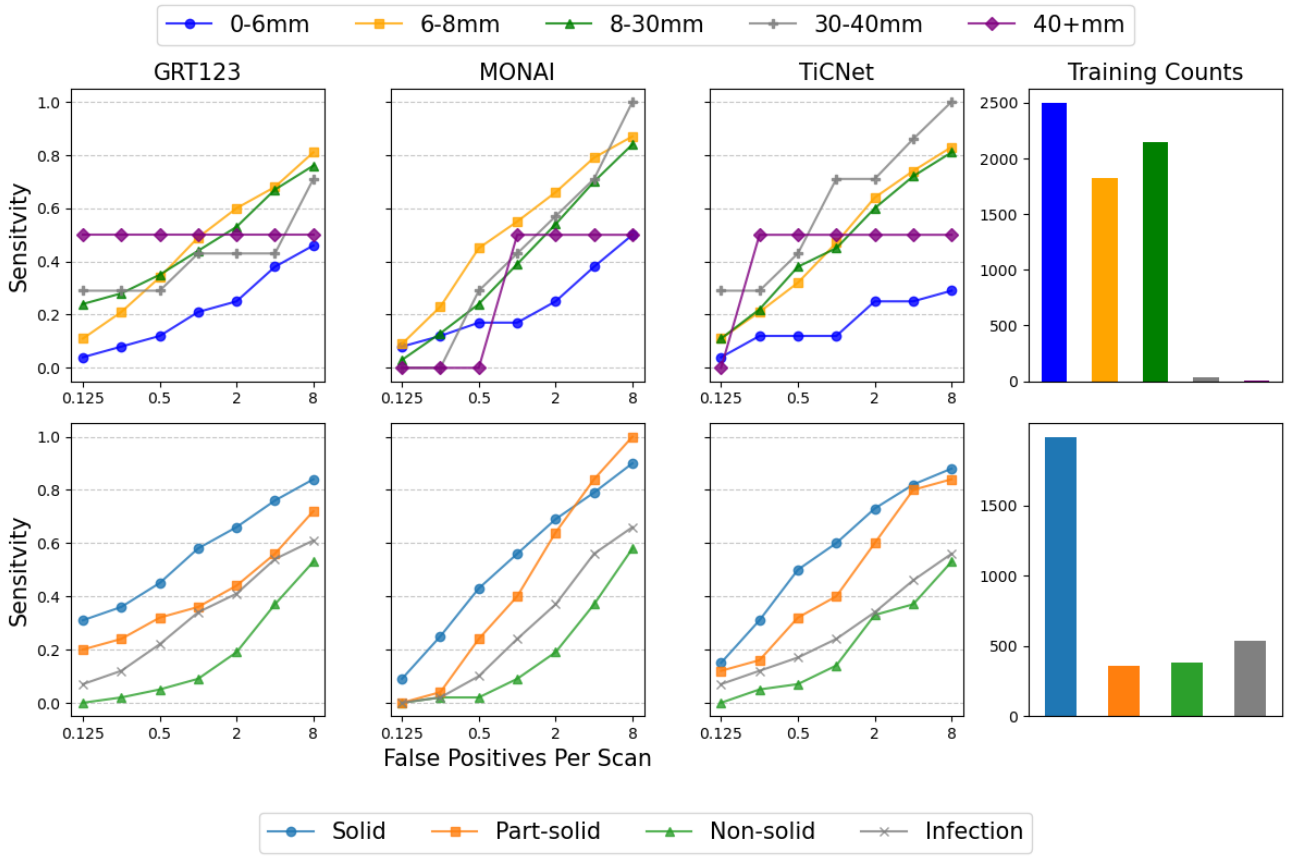

Figure 18: Sensitivity for each nodule category by size (top row) and nodule type (bottom row) for GRT123 U-Net, MONAI RetinaNet and TiCNet Transformer at the fixed operating points. Training proportions are indicated in the bar plot on the right.

per scan. Specifically, groups with a higher proportion of scans containing multiple nodules tend to exhibit higher TPR, PPV, and PR, simply because the algorithms have more opportunities to detect at least one actionable nodule per scan. Conversely, groups with a higher proportion of scans with zero nodules are more likely to show decreased TPR and PPV, and increased PR, driven by false positives.

While these differences appear to stem from sampling artefacts within the test set, they are nonetheless important. For example, in the SUMMIT dataset, Black participants had significantly fewer scans with multiple nodules, which may adversely impact scan-level metrics and, more importantly, could have negative, downstream clinical implications.

At the nodule level, our results indicated no adverse impact of demographic imbalances on detection ability after accounting for confounding variables. This was further supported by evaluation on an external validation dataset, where the dominant groups in the training set, namely male and White participants, did not outperform others. In fact, underrepresented ethnic groups showed higher detection rates than their White counterparts. Across the results, some group-wise differences were statistically significant, and post-hoc analyses provided insights into the contributing factors behind these variations.

The model architecture plays a critical role not just in overall accuracy but in the specific types of nodules it prioritizes for detection. MONAI RetinaNet achieved the best overall performance, particularly excelling in the detection of small nodules at low operating points. In contrast, GRT123 U-Net and TiCNet Transformer performed better for larger nodules at these same operating points. This distinction comes from the different model designs. GRT123 U-Net, using over-sampling for nodules sized 30-40mm and 40+mm, was effective at detecting larger nodules at lower operating points. MONAI RetinaNet, leveraging the scale robustness of the FPN, was more proficient at detecting smaller nodules. TiCNet Transformer, with its focus on attention blocks, excelled in detecting the largest nodules at low operating points but struggled with smaller ones, as its attention mechanism focuses on long-range dependencies. Given that larger malignant nodules are associated with worse outcomes, it is essential for evaluation metrics to emphasize the detection of larger nodules, especially at lower operating points.

Model architectures can influence proficiency for de-

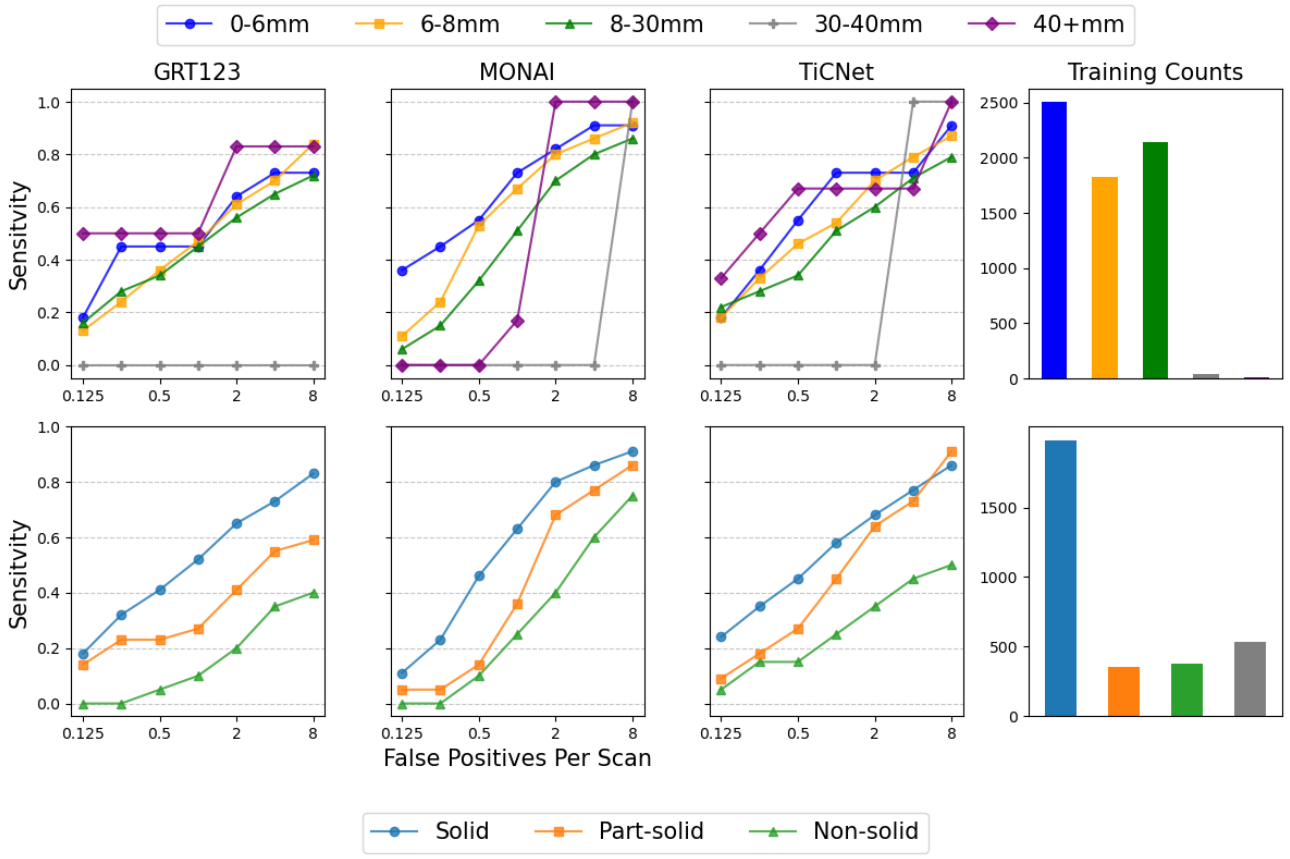

Figure 19: Sensitivity for each nodule category by size (top row) and nodule type (bottom row) for GRT123 U-Net (left), MONAI RetinaNet (middle) and TiCNet Transformer (right) at the fixed operating points. Training proportions are indicated in the bar plot on the right.

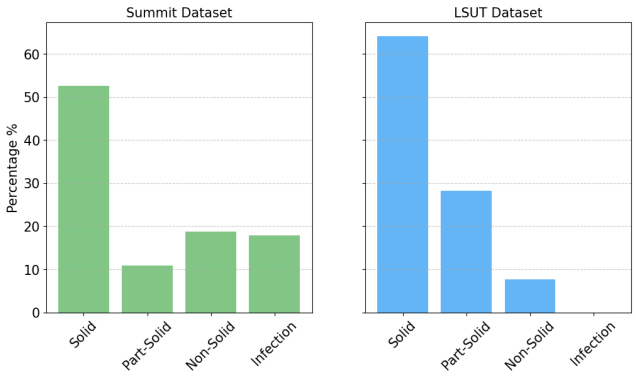

Figure 20: Proportion of actionable nodule types in the test datasets for SUMMIT and LSUT. Bars represent the percentage of each nodule type within their respective test sets. The SUMMIT test dataset is from Experiment 1.

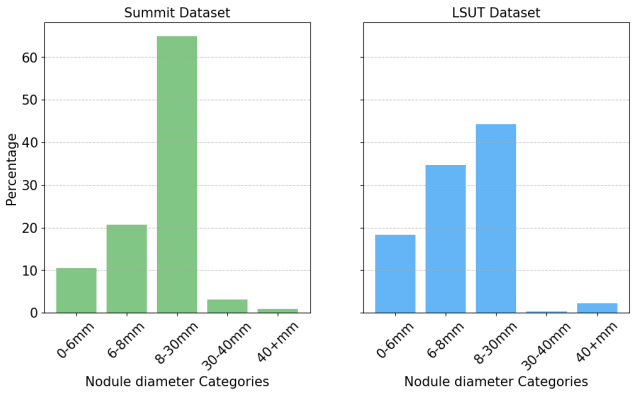

Figure 21: Proportion of actionable nodule diameters in the test datasets for SUMMIT and LSUT. Bars represent the percentage of each nodule type within their respective test sets. The SUMMIT test dataset is from Experiment 1.

tecting specific nodule types, our findings suggest that the primary determinant of performance and fairness is the composition and quality of the training data. Before optimizing architecture, it is essential to first establish which nodules and clinical features are most important to detect. Data-driven approaches that carefully curate training sets, particularly with attention to nodule subtypes and clinically actionable findings, are likely to yield greater improvements in both overall accuracy and equitable performance than relying solely on larger or more complex models.

While these interpretations are consistent with the observed distribution of detected nodule types at different FPPS rates, they should be regarded as indicative rather

than definitive. We did not conduct ablation experiments to isolate the contribution of specific architectural components, nor did we explicitly control for model capacity, which may also influence performance. Future work will be required to systematically disentangle these effects and confirm which architectural elements most strongly drive differences in the detection of nodule types and sizes. Explainability and feature attribution methods, such as SHAP or Grad-CAM, could provide additional insight into model behaviour by highlighting which imaging features drive the detection of specific nodule types or sizes. Such approaches may complement the current subtype-level analysis and guide improvements in architecture and training strategies. Nevertheless, the subtype-level analysis presented here provides useful insight into the strengths and weaknesses of each model and offers a foundation for such future investigations.

The mAP analysis revealed differences in performance across models, with GRT123 U-Net and MONAI RetinaNet experiencing only slight decreases in performance as IoU (Intersection over Union) increased. In contrast, TiCNet Transformer saw dramatic drop-offs in performance, indicating that its predicted diameters for nodules did not align with the true values.

Currently, nodule detection algorithms are optimized for the FROC curve, which prioritizes overall detection sensitivity but does not differentiate based on nodule severity or clinical management workflow. This focus may lead to models being optimized for the wrong objective, potentially overemphasizing common small, less clinically significant nodules while neglecting rarer, larger, high-risk ones. These findings highlight the need to reassess evaluation metrics and consider alternative measures that better align with clinical decision-making and downstream management plans.

The post-hoc analysis revealed that the prevalence of dominant nodule types in the training data influenced the evaluation results. Solid nodules, being the most common, were more easily detected compared to rarer nodule types. However, the overall performance did not align perfectly with the underlying training proportions, as part-solid nodules outperformed non-solid nodules. This discrepancy may be attributed to shared features between solid and part-solid nodules. As different types of lung cancer are associated with specific nodule types, and certain cancers have higher incidences in specific protected groups, it is crucial that nodule detection algorithms address the poor performance on rarer types, such as non-solid nodules.

This analysis has focused on lung cancer screening populations composed of high-risk individuals, specifically smokers. However, when considering pulmonary nodule detection algorithms for populations outside of those with a smoking history, it is important to recognize the increasing concern over lung cancer incidence among individuals who have never smoked, particularly in East Asia. In this region, lung cancer is the leading cause of cancer-related deaths, with 53% of cases occurring in never-smokers (Yang (2021)). These cases often involve part-solid nodules, which, despite performing adequately in our study due to shared features with solid nodules, showed significantly lower detection performance compared to solid nodules. As such, applying a nodule detection algorithm trained on a lung cancer screening dataset from smokers to a population of never-smokers may not yield equivalent performance, potentially leading to poorer patient outcomes. This highlights the need for developing algorithms that can incorporate our increasingly granular understanding of individual risk, accounting for diverse risk factors and nodule characteristics.

This analysis also challenges the one-size-fits-all approaches in nodule detection algorithms. Further research into developing algorithms tailored to specific nodule types and sizes could enhance both their explainability and generalizability.

Our study also highlights the importance of understanding the key drivers of algorithmic performance, emphasizing the need for algorithms to clearly specify their training conditions and areas of proficiency and limitation. Further research and validation are essential to ensure that these algorithms maintain their effectiveness and fairness when introduced to new populations and clinical settings.

The limitations of this study primarily stem from the characteristics and design of the SUMMIT and LSUT datasets, as well as the performance and annotation practices of the evaluated models. Both cohorts were sampled exclusively from the London area, which restricts the generalisability of findings to other geographical regions, and all SUMMIT scans were acquired using a uniform protocol and scanner type, differing from the variability encountered in routine clinical practice. The LSUT validation cohort, while providing valuable external confirmation of overall model performance, is relatively small and includes limited representation of non-white groups, constraining the scope of subgroup fairness analyses. Annotation strategies further limit the granularity of analyses, in SUMMIT and LSUT, radiologists prioritized identifying actionable nodules for clinical management rather than exhaustively cataloguing all benign findings. This real-world workflow contrasts with benchmarking datasets such as LUNA16, where comprehensive annotation allows more detailed subtype-level evaluation.

While the models and datasets used in this study are not identical to those deployed in commercial CADe systems, the findings provide valuable insights for understanding model behaviour in real-world screening scenarios. Our results indicate that, within the constraints of currently available training data, unbalanced demographic distributions do not necessarily lead to systematic disparities in nodule detection. Instead, performance is largely driven by

the composition of the training dataset and the distribution of nodule types. This highlights the importance of carefully curated, modern benchmarking datasets, ideally annotated down to nodule subtype, for both commercial evaluation and regulatory assessment. Such datasets can help identify gaps in algorithm performance, support fairness evaluations, and guide improvements before deployment. While our study does not establish regulatory standards, it illustrates how a structured evaluation framework combining demographic and subtype-level analyses can be applied to ensure robust, equitable, and clinically meaningful performance of pulmonary nodule detection algorithms.

The evaluated models also demonstrated lower overall CPM compared to top-performing LUNA16 entries, which increases the risk of both false-positive and false-negative conclusions in fairness analyses. Contributing factors include differences in reference standards (e.g., qualifying diameters), broader case heterogeneity in real-world cohorts, and the absence of comprehensive benign-nodule annotation in clinically focused datasets. Many high-ranking LUNA16 models are not reproducible due to unavailable code, whereas the architectures used here are openly available and representative of current practice. Taken together, these factors highlight that performance and fairness conclusions are constrained by dataset composition, demographic representation, nodule subtype distribution and baseline algorithmic performance.

Future studies could overcome these limitations by incorporating datasets with diverse scanning protocols and broader geographical representation, increasing variability in both imaging conditions and demographics. Extending the scope of nodule detection algorithms would also provide more insight into how design choices impact performance across different nodule types and sizes. Furthermore, enhancing annotation completeness, by focusing on actionable and benign nodules, would enable more robust fairness and subtype-level analyses. Finally, refining evaluation metrics to better reflect clinical impact would ensure that algorithm performance assessments are meaningful for real-world screening scenarios.

This study also revealed that across all studied models, calibration was inadequate, with the models being generally over-confident, a finding that was consistent across studied sub-groups. To be able to meaningfully use the output probabilities in a clinical workflow, efforts should be exerted first to understand the causes of such poor calibration, alleviate it through recalibration methods or changes in model architecture and assess generalisability of the solution in multiple populations.

Our findings fit within a growing body of work on fairness in pulmonary nodule detection. Studies have shown that algorithm performance can vary across demographic groups and nodule characteristics, often reflecting biases in training data (Sourlos et al. (2022)). Ensuring equitable performance requires diverse, well-annotated datasets and validation across populations (Jia et al. (2025)). Public resources such as LUNA16 and fine-grained pulmonary lesion datasets are critical for benchmarking and model development, though gaps remain, particularly in longitudinal imaging and tracking tumor evolution (Mei et al. (2021)). Together, these works emphasize the need for demographic-aware evaluation and careful dataset design to mitigate bias and improve fairness in AI-based nodule detection.

In conclusion, our analysis demonstrates reassuring results regarding the overall fairness of existing nodule detection algorithms. However, it also underscores the need for continued efforts to ensure sustained performance not only across diverse demographic populations but also nodule types. It challenges the one-size-fits-all approaches in nodule detection algorithms and promotes the need for the development of solutions tailored to discrete nodule characteristics.

Acknowledgments

We would like to acknowledge the contributions of the following:

SUMMIT Consortium

Jennifer L Dickson[1], Carolyn Horst[1], Sophie Tisi[1], Helen Hall[1], Priyam Verghese[1], Andrew Creamer[1], Ruth Prendecki[1], Anne-Marie Hacker[2], Anand Devaraj[5,6], Kylie Gyertson[4], Neal Navani[4], Anant Patel[7], Alan Shaw[8], Rajesh Banka[9], Angshu Bhowmik[10], Ugo Ekeowa[11], Chris Valerio[12], William M Ricketts[13], Aashna Samson[4], Adefiola Olabode[4], Alberto Villanueva[4], Ali Mohammed[13], Alice Cotton[4], Andrew Perugia[22], Anjeli Ketkar[4], Anna Sikorska[4], Anthony Edey[4], Antonette Andrews[4], April Neville[4], Burcu Ozaltin[20], Camilla Mitchell[4], Catherine Nestor[4], Celia Snell[4], Charlie Sayer[4], Charlotte Cash[7], Chimtom Nwaosu[4], Christine Hosein[4], Claire Levermore[4], Columbus Ife[4], Derya Ovayolu[4], Dominique Arancon[4], Domminique Arancon[4], Eleanor Hellier[4], Elena Stefan[4], Elodie Murali[4], Esther Arthur-Darkwa[4], Ethaar El-Emir[4], Fabia Ruggiero[4], Fahiza Chowdhury[4], Fahmida Hoque[4], Fanta Bojang[4], Farhida Begum[4], Geoff Bellingan[4], Georgia Bullock[4], Graham Robinson[4], Gulistan Keskinbicak[4], Gulten Geneci[4], Harriet Sorrell[4], Hasti Robbie[4], Helen Kiconco[4], Helen Victoria Bowyer[4], Henna Makwana[4], Hina Pervez[4], Houda Benmebarek[4], James Rusius[22], Jane Rowlands[4], Janine Zylstra[4], Jeannie Eng[4], Jhanara Begum[4], Jon Teague[2], Joseph Jacob[20], Judy Airebamen[4], Julian McKee[4], Karen Parry-Billings[4], Karen Sennett[23], Kate Davies[1], Kate Gowers[1], Kaylene Phua[4], Kim Sorley[4], Kitty Chan[2], Kylie Gyertson[4], Laila Ahmed[4], Laura Farrelly[2], Laura Green[4], Lucas Clarke[4], Lynsey Gallagher[4], Magali Taylor[4], Malavika Suresh[2], Mamta

Ruparel[1], Marie Calligaris[4], Marilyn Ijeomah-Orji[4], Mark Clark[4], Maureen Browne[4], Mehran Azimbagirad[20], Miguel Ferreira[4], Milica Rajkov[4], Moksud Miah[4], Monika Sommerrey[4], Navinah Nundlall[4], Nazim Miah[4], Neil Songsong[4], Nicholas Beech[4], Nick Screaton[19], Nikita Sharma[4], Olufunmilola Ajayi[4], Oluwadara Akerewusi[4], Patricia Castello[4], Paul Hickey[4], Paul Robinson[4], Qin Neville[4], Rabiya Patel[4], Rachael Sarpong[2], Ricardo McEwen[4], Rosie Dapaah Tweneboah[4], Rowan Faiers[4], Sahana Uthayakumar[4], Sam Hare[4], Samanjit Hare[7], Sara Lock[8], Sebri Hasaballa[4], Shaira Hassan[4], Sheetal Karavadra[4], Shummi Begum[4], Simranjit Mehta[4], Sofia Nnorom[4], Stephen Ellis[4], Stuart Marnham[4], Sumiya Mahey[4], Suraiya Sharmin[4], Syeda Anam Hussain[4], Sylvia Patricia Enes[4], Tania Anastasiadis[21], Tania Anastasiadis[21], Tanita Limani[4], Terry O'Shaughnessy[13], Thea Buchan[4], Theepan Visakan[4], Thomas Callender[1], Tracy Odigie[4], Trisha Enes[4], Tunku Aziz[13], Urja Patel[4], Vicky Bowyer[4], Ya-Fei Li[4], Yee Chin Lee[4], Zaheer Mangera[12], Zahra Awan[4], Zahra Hanif[4], Zelena Aziz[4],

Affliations

[1] Lungs for Living Research Centre, UCL Respiratory, University College London, London [2] CRUK UCL Cancer Trials Centre, University College London, London [3] Centre for Cancer Screening, Prevention, Detection and Early Diagnosis, Wolfson Institute of Population Health, Barts and The London School of Medicine and Dentistry, Queen Mary University of London, London [4] University College London Hospitals NHS Foundation Trust, London [5,6] RBHT, Imperial College [7] Royal Free London NHS Foundation Trust, London [8] Whittington Health NHS Trust, London [9] Barking, Havering and Redbridge University Hospitals NHS Trust, Essex [10] Homerton University Hospital Foundation Trust, London [11] The Princess Alexandra Hospital NHS Trust, Essex [12] North Middlesex University Hospital NHS Trust, London [13] Barts Health NHS Trust, London [19] Royal Papworth Hospital NHS Foundation Trust, Cambridge [20] Satsuma Lab, Hawkes Institute, London [21] Tower Hamlets Clinical Commissioning Group, London [22] Noclor Research Support, London [23] Islington Clinical Commissioning Group, London

Ethical Standards

The work follows appropriate ethical standards in conducting research and writing the manuscript, following all applicable laws and regulations regarding treatment of animals or human subjects.

Conflicts of Interest

JM is an investigator for the SUMMIT study within which the present study cohort is embedded. SUMMIT is sponsored and conducted by University College London and funded by GRAIL LLP through a research grant awarded to SMJ as principal investigator.

The SUMMIT Study is funded by GRAIL LLC. through a research grant awarded to SMJ as Principal Investigator. SMJ is supported by CRUK programme grant (EDD-CPGM\100002), and MRC Programme grant (MR/W025051/1). SMJ receives support from the CRUK Lung Cancer Centre of Excellence (C11496/A30025) and the CRUK City of London Centre, the Rosetrees Trust, the Roy Castle Lung Cancer Foundation, the Longfonds BREATH Consortia, MRC UKRMP2 Consortia, the Garfield Weston Trust, and University College London Hospitals Charitable Foundation.

SMJ's work is supported by a Stand Up To Cancer-LUNGevity-American Lung Association Lung Cancer Interception Dream Team Translational Research Grant and Johnson and Johnson (grant number: SU2C-AACR-DT23-17 to S.M. Dubinett and A.E. Spira). Stand Up To Cancer is a division of the Entertainment Industry Foundation. Research grants are administered by the American Association for Cancer Research, the Scientific Partner of SU2C. This work was partly undertaken at UCLH/UCL, which received a proportion of funding from the Department of Health's NIHR Biomedical Research Centre's funding scheme.

SMJ has received fees for advisory board membership in the last three years from Bard1 Lifescience. He has received grant income from GRAIL Inc. He is an unpaid member of a GRAIL advisory board. He has received lecture fees for academic meetings from Cheisi and Astra Zeneca. His wife works for Astra Zeneca.

This research was funded in whole or in part by the Wellcome Trust (227835/Z/23/Z). J Jacob was also supported by the NIHR UCLH Biomedical Research Centre.

JJ declares consultancy fees from Boehringer Ingelheim, F. Hoffmann-La Roche, GlaxoSmithKline, NHSX; fees from advisory Boards for Boehringer Ingelheim, F. Hoffmann-La Roche; lecture fees from Boehringer Ingelheim, F. Hoffmann-La Roche, Takeda; grant funding from GlaxoSmithKline, Wellcome Trust, Microsoft Research, Gilead Sciences, Chan Zuckerberg Initiative and UK patent application numbers: 2113765.8 and GB2211487.0.

TP receives funding from AstraZeneca for a Clinical Research Fellowship.

MM recieves funding from the University of British Columbia through the clinician investigator program and has received travel funding from GRAIL LLP.

AN is a member of the advisory board for Aidence BV and Faculty Sciences Ltd.

AH has received an honorarium for an advisory bord meeting for GRAIL, a consultation fee for Evidera Inc for a GRAIL initiated project, and previously owned shares in Illumina.

Data availability

Any requests for data sharing for SUMMIT data must be sent to the Sponsor (via ctc.summit@ucl.ac.uk), who will be able to deal with this request. Data requests must include the following: data being requested, purpose of data request, your name, contact details and institution/organisation. Please be aware data may not be immediately available due to processing, this data will be made available in time.

References

H. Arnold. Paving the way for lung cancer screening: Learning from central and eastern europe, 2024. URL https://www.lungcancerpolicynetwork.com. Accessed: 2025-03-19.

A.A. Bankier, H. MacMahon, T. Colby, P.A. Gevenois, J.M. Goo, A.N.C. Leung, D.A. Lynch, C.M. Schaefer-Prokop, N. Tomiyama, W.D. Travis, J.A. Verschakelen, C.S. White, and D.P. Naidich. Fleischner society: Glossary of terms for thoracic imaging. *Radiology*, 310: e232558, 2024. .

S. Beyaz. A brief history of artificial intelligence and robotic surgery in orthopedics & traumatology and future expectations. *Joint Diseases and Related Surgery*, 31:653–655, 2020. .

M. Brandao. Age and gender bias in pedestrian detection algorithms, 2019. URL http://arxiv.org/abs/1906.10490.

F. Bray, M. Laversanne, H. Sung, J. Ferlay, R.L. Siegel, I. Soerjomataram, and A. Jemal. Global cancer statistics 2022: Globocan estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA: A Cancer Journal for Clinicians*, 74:229–263, 2024. . URL https://doi.org/10.3322/caac.21834.

British Thoracic Society. Pulmonary nodules: British thoracic society guidelines, 2015. URL https://www.brit-thoracic.org.uk/quality-improvement/guidelines/pulmonary-nodules/. Accessed: 2025-06-06.

M. Bucher, S. Herbin, and F. Jurie. Hard negative mining for metric learning based zero-shot classification, 2016. URL http://arxiv.org/abs/1608.07441.

J. Buolamwini and T. Gebru. Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Proceedings of the 1st Conference on Fairness, Accountability and Transparency*, pages 77–91. PMLR, 2018.

P. Burlina, N. Joshi, W. Paul, K.D. Pacheco, and N.M. Bressler. Addressing artificial intelligence bias in retinal diagnostics. *Translational Vision Science & Technology*, 10:13, 2021. .

Cancer Research UK. Cancer statistics for the uk, 2015. URL https://www.cancerresearchuk.org/health-professional/cancer-statistics-for-the-uk. Accessed: 2023-03-23.

D.T.-H. Chen, J. Hirst, C.A.C. Coupland, W. Liao, D.R. Baldwin, and J. Hippisley-Cox. Ethnic disparities in lung cancer incidence and differences in diagnostic characteristics: a population-based cohort study in england. *The Lancet Regional Health – Europe*, 48, 2025. .

Hong Chen, Haozhe Huang, Jianye Zhang, Xuexue Wang, Mengyang Han, Chanjun Ding, and Jinhong Wang. Intra- and inter-reader variations in lung nodule measurements: Influences of nodule size, location, and observers. *Diagnostics*, 12(10):2319, 2022. . URL https://www.mdpi.com/2075-4418/12/10/2319.

H.J. de Koning, C.M. van der Aalst, P.A. de Jong, E.T. Scholten, K. Nackaerts, M.A. Heuvelmans, and et al. Reduced lung-cancer mortality with volume ct screening in a randomized trial. *New England Journal of Medicine*, 382:503–513, 2020. .

A. Gyrard, S. Abedian, P. Gribbon, G. Manias, R. van Nuland, K. Zatloukal, I.E. Nicolae, G. Danciu, S. Nechifor, L. Marti-Bonmati, P. Mallol, S. Dalmiani, S. Autexier, M. Jendrossek, I. Avramidis, E.G. Alvarez, P. Holub, I. Blanquer, A. Boden, and R. Hussein. Lessons learned from european health data projects with cancer use cases: Implementation of health standards and internet of things semantic interoperability. *Journal of Medical Internet Research*, 27:e66273, 2025. .

T.N. Hanna, S.D. Steenburg, A.B. Rosenkrantz, R.S. Pyatt, R. Duszak, and E.B. Friedberg. Emerging challenges and opportunities in the evolution of teleradiology. *American Journal of Roentgenology*, 215:1411–1416, 2020. .

Ruiting Jia, Baozhi Liu, and Mohsin Ali. Establishing an ai-based diagnostic framework for pulmonary nodules in computed tomography. *BMC Pulmonary Medicine*, 25(1):339, 2025. . URL https://doi.org/10.1186/s12890-025-03806-7.

Kaggle. Data science bowl 2017, 2017. URL https://kaggle.com/data-science-bowl-2017. last accessed 2025/03/19.

F. Kuppers, J. Kronenberger, A. Shantia, and A. Haselhoff. Multivariate confidence calibration for object detection. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, page 1322–1330, Seattle, WA, USA, 2020. IEEE. .

F. Liao, M. Liang, Z. Li, X. Hu, and S. Song. Evaluate the malignancy of pulmonary nodules using the 3d deep leaky noisy-or network. *IEEE Trans. Neural Netw. Learning Syst.*, 30:3484–3495, 2019. .

T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie. Feature pyramid networks for object detection, 2017. URL http://arxiv.org/abs/1612.03144.

T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár. Focal loss for dense object detection, 2018a. URL http://arxiv.org/abs/1708.02002.

T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár. Focal loss for dense object detection, 2018b. URL http://arxiv.org/abs/1708.02002.

Ling Ma, Gen Li, Xingyu Feng, Qiliang Fan, and Lizhi Liu. Ticnet: Transformer in convolutional neural network for pulmonary nodule detection on ct images. *Journal of Imaging Informatics in Medicine*, pages 1–13, 2024.

J. McCabe, D. Cheng, A. Bhamani, M. Mullin, T. Patrick, A. Nair, S.M. Janes, C.H. Sudre, and J. Jacob. Exploring fairness in state-of-the-art pulmonary nodule detection algorithms, 2024. URL https://doi.org/dx.doi.org/10.1007/978-3-031-72787-0_8. last accessed 2025/03/19.

Jie Mei, Ming-Ming Cheng, Gang Xu, Lan-Ruo Wan, and Huan Zhang. Sanet: A slice-aware network for pulmonary nodule detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 1–1, 2021. . URL https://mmcheng.net/SANet/.

S.A. Mirak, S.H. Tirumani, N. Ramaiya, and I. Mohamed. The growing nationwide radiologist shortage: Current opportunities and ongoing challenges for international medical graduate radiologists. *Radiology*, 2025. .

National Lung Screening Trial Research Team. The national lung screening trial: Overview and study design. *Radiology*, 258:243, 2011. .

J.R. Parikh, D. Wolfman, C.E. Bender, and E. Arleo. Radiologist burnout according to surveyed radiology practice leaders. *Journal of the American College of Radiology*, 17:78–81, 2020. .

E. Puyol-Anton, B. Ruijsink, S.K. Piechnik, S. Neubauer, S.E. Petersen, R. Razavi, and A.P. King. Fairness in cardiac mr image analysis: An investigation of bias due to data imbalance in deep learning based segmentation, 2021. URL http://arxiv.org/abs/2106.12387.

O. Ronneberger, P. Fischer, and T. Brox. U-net: Convolutional networks for biomedical image segmentation, 2015. URL http://arxiv.org/abs/1505.04597.

E. Roy, J. Shrager, J. Benson, W.L. Trope, P. Bhandari, N. Lui, D. Liou, L. Backhus, and M.F. Berry. Risk of adenocarcinoma in patients with a suspicious ground-glass opacity: a retrospective review. *J Thorac Dis.*, 14:4236–4245, 2022. .

A.A.A. Setio, A. Traverso, T. de Bel, M.S.N. Berens, C. van den Bogaard, P. Cerello, H. Chen, Q. Dou, M.E. Fantacci, B. Geurts, R. van der Gugten, P.A. Heng, B. Jansen, M.M.J. de Kaste, V. Kotov, J.Y.-H. Lin, J.T.M.C. Manders, A. Sóñora-Mengana, J.C. García-Naranjo, E. Papavasileiou, M. Prokop, M. Saletta, C.M. Schaefer-Prokop, E.T. Scholten, L. Scholten, M.M. Snoeren, E.L. Torres, J. Vandemeulebroucke, N. Walasek, G.C.A. Zuidhof, B. van Ginneken, and C. Jacobs. Validation, comparison, and combination of algorithms for automatic detection of pulmonary nodules in computed tomography images: The luna16 challenge. *Medical Image Analysis*, 42:1–13, 2017. . URL https://doi.org/10.1016/j.media.2017.06.015.

S. Singh, J. Maxwell, J.A. Baker, J.L. Nicholas, and J.Y. Lo. Computer-aided classification of breast masses: Performance and interobserver variability of expert radiologists versus residents. *Radiology*, 258(1):73–80, 2011. .

Nikos Sourlos, Jingxuan Wang, Yeshaswini Nagaraj, Peter van Ooijen, and Rozemarijn Vliegenthart. Possible bias in supervised deep learning algorithms for ct lung nodule detection and classification. *Cancers*, 14(16):3867, 2022. . URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC9405732/.

The Royal College of Radiologists. Rcr clinical radiology census report 2021, 2021. URL https://www.rcr.ac.uk/clinical-radiology/rcr-clinical-radiology-census-report-2021. Accessed: 2023-03-23.

M. Wang and W. Deng. Mitigate bias in face recognition using skewness-aware reinforcement learning, 2019. URL http://arxiv.org/abs/1911.10692.

N. Weng, S. Bigdeli, E. Petersen, and A. Feragen. Are sex-based physiological differences the cause of gender bias for chest x-ray diagnosis?, 2023.

P. Yang. Ps01.02 national lung cancer screening program in taiwan: The talent study. *Journal of Thoracic Oncology*, 16:S58, 2021. .

X. Yang, Y. Wang, R. Byrne, G. Schneider, and S. Yang. Concepts of artificial intelligence for computer-assisted drug discovery. *Chemical Reviews*, 119:10520–10594, 2019. .