

Repurposing adult AI radiograph models for neonatal care through continual learning: an international multicentre study

Received: 22 December 2025

Accepted: 8 September 2026

Published online: 13 October 2026

 Check for updates

A list of authors and their affiliations appears at the end of the paper

Despite growing clinical adoption of artificial intelligence in radiology, fewer than one in five commercially available models are cleared for paediatric use. Existing systems are trained predominantly on adult data and generalize poorly to neonates, yet efficient strategies for repurposing them without data sharing or large-scale local data curation are lacking. Here we present an international retrospective study of neonatal chest radiographs from 30 neonatal intensive care units across 20 countries and territories, evaluating continual learning as a privacy-preserving approach to adapt a deep learning model for adult endotracheal tube assessment to neonatal care. Each hospital contributed 50 radiographs for sequential model retraining without centralized data pooling. Continual learning achieved lower anatomical landmark localization error (0.67 cm versus 1.29 cm for hospital-specific fine-tuning and 4.30 cm for adult-only models) at 29 of 30 sites, with improved detection of deep intubations at 27 of 30 sites. These findings demonstrate that privacy-preserving collaborative learning can address critical gaps in paediatric AI deployment.

Current medical artificial intelligence (AI) models predominantly focus on adults, and only 17% of commercial models have received US Food and Drug Administration (FDA) approval for use in paediatric patients^{1,2}. Within radiology, AI tools intended for adults may be difficult to directly apply to neonatal patients, as differences in physiology and imaging protocols can cause substantial drops in performance^{3,4}. While some studies have attempted to adapt models originally trained on adult data to neonates, this research has historically been limited by a lack of neonatal data^{5,6}. Assembling large, annotated paediatric imaging datasets is particularly challenging: curating and labelling clinical data requires substantial effort, data sharing across institutions is constrained by privacy regulations, and children represent less than 1% of public medical imaging datasets^{5,7}. As a consequence, the majority of paediatric AI models rely on single-centre data, with few reporting external validation and many experiencing substantial performance degradation in cross-institutional settings^{8,9}. These barriers underscore the need for data-efficient adaptation techniques that can leverage existing adult AI models for paediatric settings and validate them across diverse environments.

To address these challenges, we conducted an international retrospective study across 30 neonatal intensive care units (NICUs) in 20 countries and territories to validate a collaborative, multihospital approach for adapting an adult AI model to neonatal patients using only small amounts of locally curated data from each site, without requiring any institution to share patient data. We focused on endotracheal tube (ETT) assessment as a proof-of-concept clinical application, as it is a task for which multiple AI tools have received FDA clearance for adult radiographs^{10,11}, yet no commercial solutions exist for neonatal populations. While some paediatric studies have used clinical and demographic variables to predict ETT size or inappropriate ETT depth^{12,13}, no previous work has performed radiograph-based ETT screening in neonates. Commonly used in both intensive care unit (ICU) and NICU settings, ETTs are crucial for treating respiratory issues, yet correct placement of ETTs in neonates can be a challenging task with little margin for error¹⁴. Identifying incorrectly placed tubes is time sensitive, and if left unaddressed, abnormal intubations can lead to serious complications. Deep intubations risk single lung ventilation and lung collapse¹⁵, while false alarms may prompt unnecessary re-intubation attempts

✉ e-mail: pranav_rajpurkar@hms.harvard.edu

that increase procedural risk and haemodynamic instability^{15,16}. Some evidence suggests that females and younger children are especially susceptible to low intubations¹⁷.

An AI model capable of flagging concerning radiographs at the bedside and triaging them for radiologist review could shorten the time to radiographic confirmation for the most critical cases. Clinicians often interpret radiographs during initial assessment, and incorrect placements may be missed until further radiology review^{17–19}. Previous work suggests that less experienced interpreters are more likely to misidentify the position of tubes, and even the most experienced readers may occasionally overlook inaccurate placements²⁰. Furthermore, annotating the anatomical landmarks relevant to classifying abnormal intubations could assist health providers in determining the adequacy of placement. ETT screening models have already demonstrated efficacy in this clinical setting^{21,22}.

We hypothesized that continual learning could effectively and efficiently adapt an adult ETT assessment model to neonatal settings. Continual learning is a machine learning paradigm in which a model sequentially learns from a stream of data, without access to previously seen data, with the objective of accumulating and retaining knowledge across all encountered tasks^{23–25}. This paradigm is well suited to our task, where a model can be sequentially exposed to diverse neonatal data, progressively building a more generalizable representation without requiring centralized data pooling. Several related approaches for privacy-preserving multi-institutional collaboration have been proposed. Federated learning is an aggregation-based approach in which a central server coordinates training by collecting and averaging parameter updates from multiple sites across iterative communication rounds^{26–28}. Serial weight transfer, also referred to as the ‘travelling model’ paradigm, is a sequential approach in which a model is passed between institutions, with each site fine-tuning the model on its local data before forwarding the updated weights to the next site; variants include single-pass transfer and cyclical transfer, where the model is passed through participating sites multiple times²⁹. Our approach is methodologically based on single-pass serial weight transfer, but we frame and evaluate it as continual learning: in addition to assessing overall model performance, we explicitly evaluate whether the model retains and accumulates knowledge by assessing whether the final model retains or improves performance on every previously seen hospital³⁰.

Concretely, in our approach, a model would be passed from hospital to hospital, and each hospital would retrain the model on a small amount of data within its own information technology (IT) environment. Thus, hospitals can collaboratively adapt a model to NICU settings in a straightforward, data-efficient manner. Importantly, this approach does not require any hospital to share their private data, curate a large dataset, or use sophisticated distributed learning techniques. To test our hypothesis, we evaluated models across 30 NICUs in 20 countries and territories on two objectives: anatomical landmark localization accuracy of the ETT tip and carina, and downstream triaging of abnormal intubations, and we compared continual learning to conventional fine-tuning and direct prediction approaches. These baselines represent the realistic alternatives currently available to individual hospitals: applying the adult model without adaptation (direct prediction) or adapting it using only local data (fine-tuning). We focused on detecting deep intubations given their clinical severity, higher prevalence^{31,32} and relatively clear correspondence with shorter ETT–carina distance. Ultimately, we demonstrate a proof of concept for continual learning as an effective data-efficient approach for adapting adult AI image interpretation models to paediatric settings, and we systematically characterize the conditions under which this paradigm succeeds, including the role of adult pretraining, scaling behaviour with the number of contributing sites, and sensitivity to data quality, providing an evidence base applicable to multisite model adaptation beyond this specific clinical task. An overview of our approach is outlined in Fig. 1.

Results

Continual learning for paediatric radiograph interpretation

To evaluate whether adult AI models could be efficiently adapted for neonatal care through multihospital collaboration, we established an international consortium of 30 NICUs across 20 countries and territories, each contributing ~100 chest radiographs with expert annotations of endotracheal tube tip and carina positions (Fig. 1a). We implemented a continual learning framework in which a pre-existing model trained on adult ICU radiographs was sequentially passed between hospitals, with each site performing local retraining on 50 radiographs within its own computing environment before forwarding the updated model to the next institution. Concretely, the first hospital in the sequence received the adult-trained model, fine-tuned it on its local dataset, and passed the updated weights to the second hospital; this process was repeated through every site within the continual learning training sequence, with each hospital’s contribution building upon the cumulative knowledge of all preceding institutions. Because only model parameters are transferred between sites, this approach preserves data privacy by eliminating the need for centralized data pooling while enabling collaborative model adaptation across diverse clinical settings.

We compared continual learning against two baseline approaches: direct prediction, in which the original adult-trained model was applied to neonatal radiographs without adaptation, and hospital-specific fine-tuning, in which the adult model was retrained using the 50 local training radiographs from each target hospital (Fig. 1b). For each of the 30 test sites, we constructed continual learning models using data from all other hospitals in the training sequence (not including train data from the hospital being evaluated), averaging results across 10 random hospital orderings to account for potential ordering effects. Model performance was evaluated on anatomical landmark localization accuracy and the downstream clinical task of identifying deep intubations requiring expedited review. Detailed descriptions of the training approach, baseline comparisons and evaluation metrics are outlined in the Methods.

Study cohort

Our study cohort reflected the clinical diversity of global NICU populations. This international multicentre study contained 2,965 NICU chest X-rays from 2,960 unique patients across 30 participating centres, outlined in Supplementary Fig. 1 and Extended Data Table 1. The median age at radiography was 2 days (IQR, 1–11 days), indicating that most infants were imaged within the first few days of life. The mean birth weight was 2,195 grams (s.d., 1,234 grams) and mean gestational age was 32.9 weeks (s.d., 5.3 weeks), reflecting a predominantly preterm population. Male infants comprised 55.9% ($n = 1,654$) of the cohort, while 43.6% ($n = 1,291$) were female and 0.5% ($n = 15$) had unknown or unreported sex. Race/ethnicity was distributed as Asian (31.9%, $n = 943$), White (25.9%, $n = 767$), Other (3.2%, $n = 95$), Black or African American (2.6%, $n = 78$), American Indian/Alaska Native (0.2%, $n = 6$) and Pacific Islander (0.1%, $n = 3$), with the remaining patients having unreported race/ethnicity (36.1%, $n = 1,068$). Notably, both Children’s Hospital Colorado and Sidra Medicine had substantially older infants at imaging (median ages 36 and 28 days respectively) and higher birth weights (means 3,558 g and 3,263 g, respectively). Gestational age distributions were consistent across hospitals. A full summary of the demographics stratified between train and test sets is shown in Extended Data Table 2 and Supplementary Table 1. Of the 2,965 radiographs, 31 (1.0%) demonstrated bronchial intubation, defined as cases where the ETT tip bounding box was positioned entirely below the carina bounding box with no vertical overlap.

Performance of continual learning across 30 NICUs

We first evaluated whether continual learning improved anatomical landmark localization compared to baseline approaches. For each hospital, we measured the distance between model-predicted and

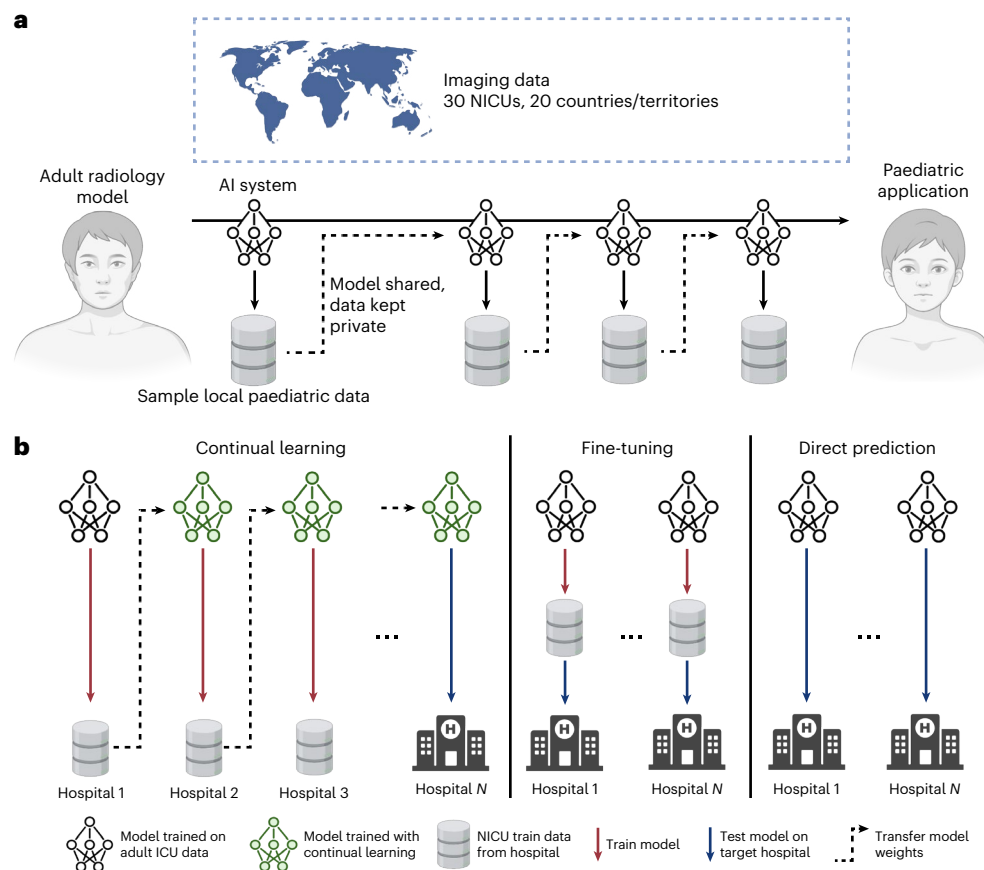


Fig. 1 | Summary of continual learning paradigm and benchmark comparisons. **a**, Overview of continual learning paradigm used to repurpose adult radiograph interpretation models to paediatric settings through a collaboration among 30 NICUs across 20 countries and territories.

b, Visualization of the methodology for continual learning compared to the fine-tuning and direct prediction baselines. N refers to the number of hospitals in the continual learning collaboration. Created in BioRender: Rao, V. <https://BioRender.com/ecupfdy> (2026).

ground-truth positions for both the ETT tip and the carina. Example predictions are shown in Fig. 2.

Continual learning models consistently outperformed both fine-tuning and direct prediction baselines across nearly all participating sites, visualized in Fig. 3. For ETT tip localization, median error decreased from 2.35 (1.59–3.10) cm using direct prediction to 0.94 (0.70–1.17) cm using fine-tuning, and further to 0.40 (0.35–0.45) cm using continual learning. Similarly, carina localization error improved from 1.62 (1.06–2.18) cm with direct prediction and 0.27 (0.25–0.30) cm with fine-tuning to 0.20 (0.19–0.22) cm with continual learning. These improvements yielded a combined tip and carina localization error of 0.67 (0.61–0.73) cm on average using continual learning, with 29 of 30 hospitals exhibiting improvement in localization error and 27 of 30 reaching statistical significance (Fig. 3a, Supplementary Table 2 and Extended Data Fig. 1).

We next assessed whether continual learning reduced clinically consequential prediction failures. We defined extreme errors as combined localization errors exceeding 5 cm that could mislead clinical decision-making. Continual learning substantially reduced the prevalence of extreme errors compared to direct prediction across all 30 hospitals, and reduced or matched fine-tuning's rate at 24 of 30 sites (Fig. 3b). On average across hospitals, >5 cm errors occurred in 2.6% of cases under continual learning, compared to 8.2% under fine-tuning and 37.4% under direct prediction.

Finally, to evaluate potential demographic disparities in model performance, we stratified localization errors by patient race, sex, age at radiography, birth weight and gestational age. Performance remained largely consistent across demographic

subgroups (Supplementary Tables 3 and 4), with no significant pairwise differences.

Clinical relevance of localization improvements

To assess whether improved anatomical localization translates to better identification of concerning ETT placements, we evaluated whether continual learning models could more effectively triage radiographs with abnormal intubations. We computed ETT–carina distances from model predictions, converting pixel distances to centimetres using the annotated ETT width as an internal calibration reference (see Supplementary Methods) and ranked radiographs by predicted distance, with shorter distances indicating concern for deep placement and longer distances indicating concern for high placements. We then evaluated the proportion of the top k concerning scans ranked by the models that contained abnormal intubations, separately for low and high intubations. We calculated precision@ k and recall@ k scores, focusing on a k value of 15 due to the distribution of the number of abnormal intubations present within each hospital's test set (Extended Data Table 1) and also reporting results for a k value of 5 in Supplementary Tables 5–7. We further measured false positive rates to assess the potential burden of unnecessary alerts.

To first validate that ground-truth ETT–carina distances are concordant with clinical placement labels, we computed Kendall's tau correlations between these distances and the clinician-assigned low/normal/high labels (Supplementary Fig. 2) and visualized the distribution of ETT–carina distances across these labels (Supplementary Fig. 3). For low intubations, the correlation was strong and significant across all 30 hospitals (mean $\tau = 0.64$, $P < 0.05$ for 30/30), confirming that

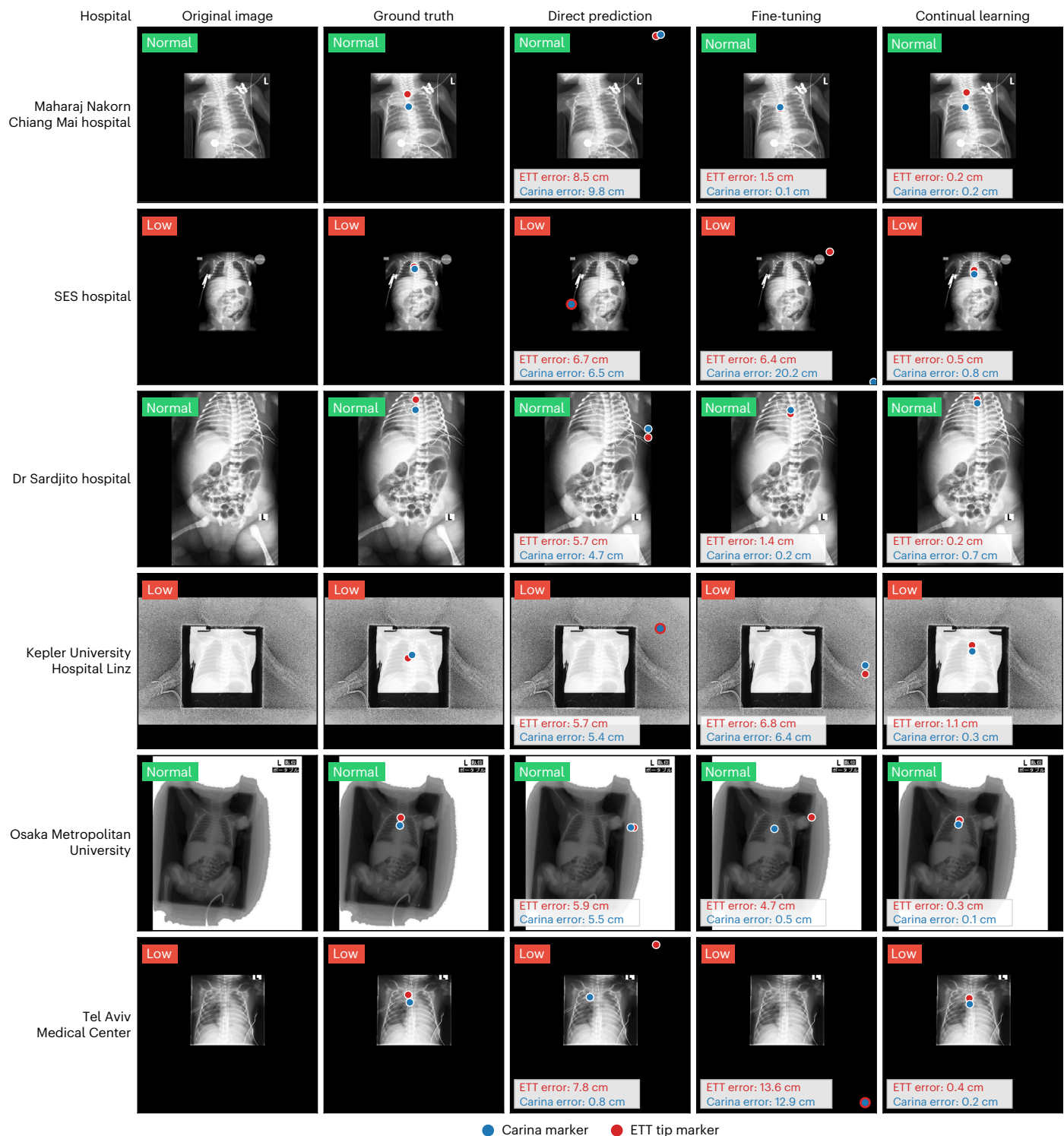


Fig. 2 | Visualization of the improvements in ETT and carina localization conferred by continual learning. Predictions displayed on sampled images across six hospitals from different countries. Rows alternate between normal and abnormally low ETT placements. Placement labels (normal or low) are

shown at the top left of each image, and prediction errors for the ETT tip and carina are shown at the bottom left of each prediction panel. Ground-truth markers represent centres of the annotated bounding boxes. Predictions are representative examples from 6 of the 30 NICU sites.

shorter ETT–carina distances reliably correspond to clinically labelled deep intubations, although lower correlations at certain sites probably reflect imprecision in the pixel-to-centimetre conversion using the ETT width. While we primarily focused on low intubations, we found that the correlation was substantially weaker for high intubations (mean $\tau = 0.39$, $P < 0.05$ for 28/29), reflecting the inherent difficulty of identifying high placements from distance measurements alone.

Continual learning models exhibited significantly higher precision@15 and recall@15 scores for detecting abnormally low ETT placements compared to fine-tuning for 18 of 30 hospitals, with positive improvements in 27 of 30 hospitals (Fig. 4). At the best-performing hospitals, continual learning models achieved precision scores over 80%. Continual learning models also achieved significantly lower false positive rates at 18 of 30 hospitals. More in-depth precision, recall and

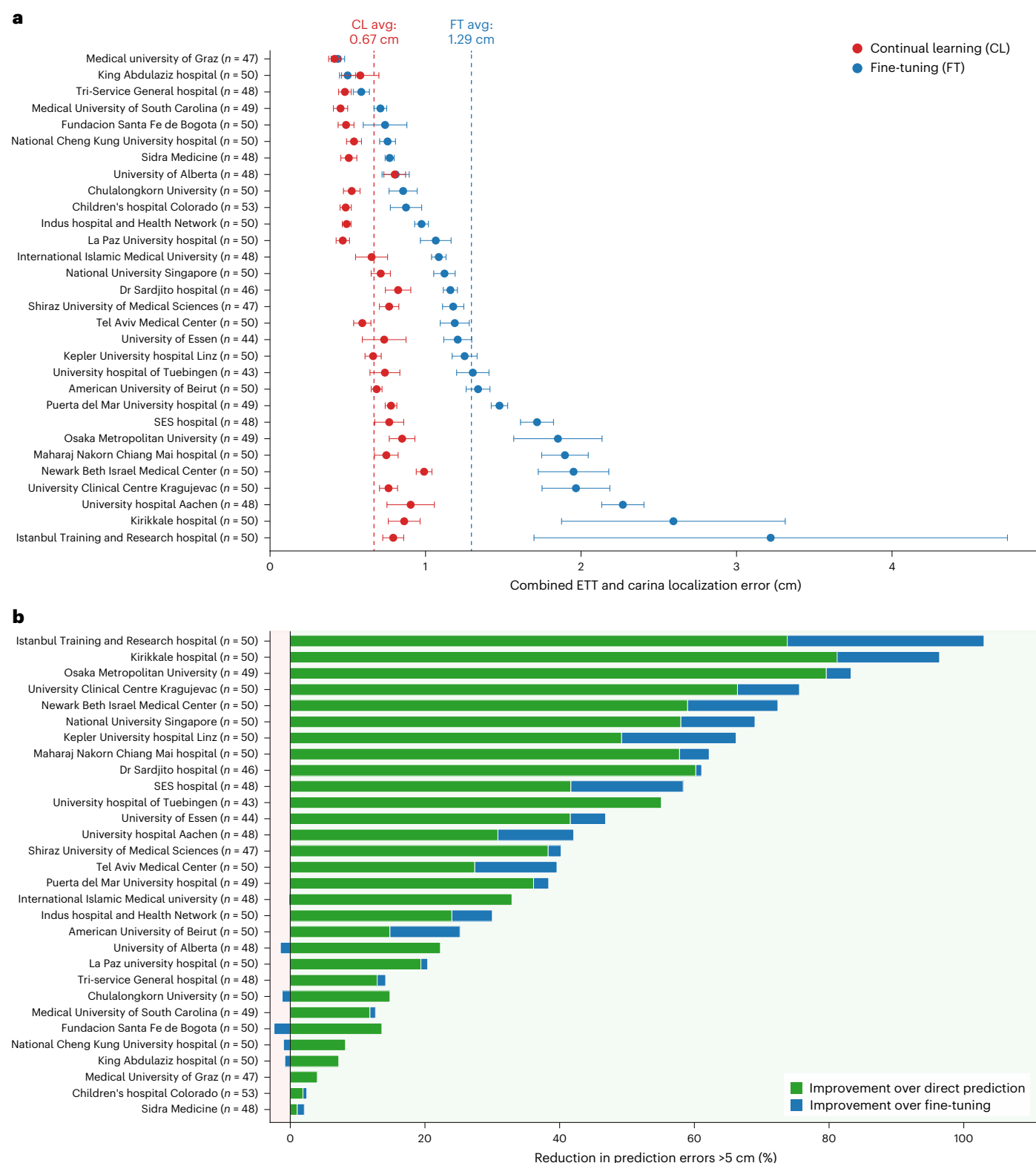


Fig. 3 | Continual learning anatomical localization performance. a, Per-site combined ETT and carina localization error for continual learning versus fine-tuning across the 30 NICU sites (each row is one site). For each method, the point shows the site's median combined ETT and carina error, averaged across 10 training simulations ($n = 10$ simulations per site; random hospital orderings

for continual learning, random training seeds for fine-tuning); error bars are 95% confidence intervals across the 10 simulations. Dashed vertical lines indicate the mean of the per-site values across all 30 sites for each method. **b**, Rate of extreme prediction errors (combined localization error >5 cm) conferred by continual learning relative to fine-tuning and direct prediction ($n = 30$ sites).

false positive rate (FPR) analysis at different k values as well as comparisons to direct prediction are reported in Supplementary Tables 5 and 6, and Extended Data Fig. 2. These findings suggest that the localization

accuracy gains conferred by continual learning could translate into clinically meaningful improvements in triaging ability for low intubations. Despite the weaker concordance between ETT–carina distances

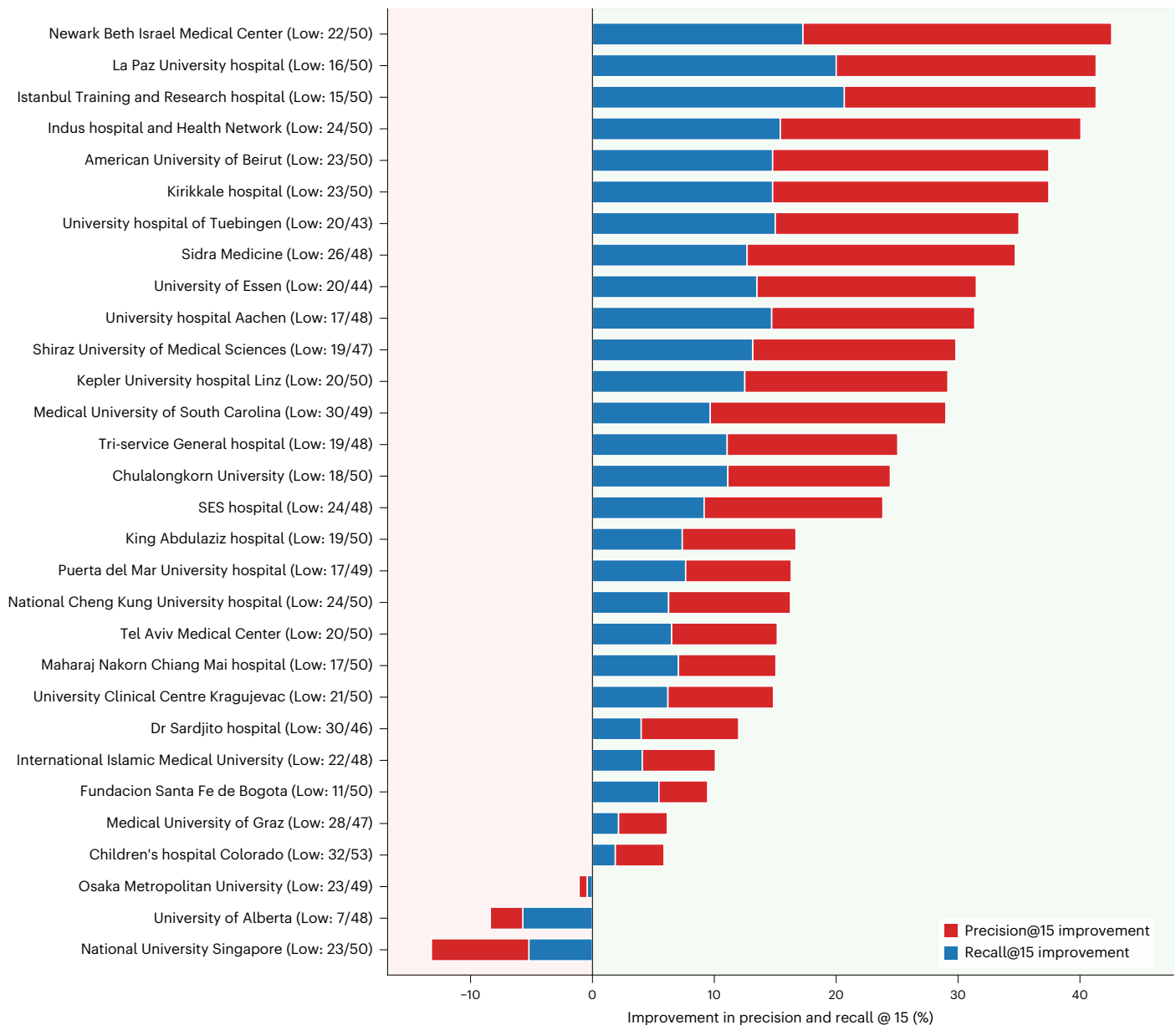


Fig. 4 | Improvement conferred by continual learning versus fine-tuning on identifying the most concerning scans for radiologist confirmation. Each bar represents 1 of the 30 NICU sites and stacks the precision@15 and recall@15 improvements for detecting abnormally low intubations. For each metric, continual learning and fine-tuning are first averaged across their 10 simulations

(random hospital orderings for continual learning, random training seeds for fine-tuning), and the bar shows the difference (continual learning–fine-tuning). Sites are ordered by total improvement. Per-site 95% confidence intervals and significance are reported in Supplementary Tables 5 and 6.

and high placement labels, continual learning models still achieved significantly higher precision@15 and recall@15 than fine-tuning at 8 of 29 hospitals (positive improvements in 23/29) for triaging high intubations, demonstrating that continual learning remains the superior adaptation strategy even for this more challenging task (Supplementary Table 7). To provide further interpretability into these results, we report the mean predicted and ground-truth ETT–carina distances for true positive and false positive cases within the top 15 flagged radiographs by continual learning models for each hospital (Supplementary Table 8).

Maintenance of performance on previously seen hospitals

To determine whether the sequential training process led to catastrophic forgetting, where performance on earlier hospitals degrades as additional sites are incorporated, we conducted a backward evaluation

comparing each hospital's performance at two timepoints: immediately after the model was trained on that hospital's data (intermediate evaluation) and after the model completed the full training sequence (final evaluation). Across 100 random orderings, the final model consistently outperformed the intermediate models at hospitals encountered early in the sequence, with a mean improvement of 0.48 cm in combined localization error at position 0 (Extended Data Fig. 3). This improvement decreased for hospitals later in the sequence (Spearman $\rho = -0.85$, $P < 0.001$), consistent with the expectation that fewer subsequent training steps provide less opportunity for additional refinement. Importantly, performance improvements remained positive or stable at nearly all sequence positions, indicating that the sequential training process does not degrade performance on previously seen hospitals but rather continues to improve it through exposure to additional institutional data. These findings provide direct evidence that the sequential

training framework supports cumulative knowledge acquisition for ETT and carina localization, suitable for broad deployment.

Extents and limits of continual learning

To better understand the factors contributing to continual learning's effectiveness, we conducted three analyses examining the roles of adult pretraining, the number of hospitals in the training sequence and the quality of radiographs contributed by participating sites.

We first evaluated the core premise underlying our approach: whether initializing the continual learning sequence with a model pretrained on adult ICU radiographs provides meaningful benefit compared to initializing with a generic object detection model without previous radiograph exposure. We hypothesized that task-relevant features learned from adult data might transfer beneficially to neonatal populations despite anatomical differences. Continual learning models initialized with adult ICU pretraining exhibited improved performance compared to models without such pretraining: anatomical localization was superior at 25 of 30 test sites, and precision@15 and recall@15 improved at 27 of 30 sites (Extended Data Fig. 4). These results support the premise that continual learning can effectively leverage adult-derived representations for paediatric application.

Having established that continual learning can effectively build upon adult-derived representations, we next investigated whether increasing the number of contributing hospitals, and thus exposure to more diverse paediatric data, would yield further improvements. Across 300 held-out-site and ordering combinations (30 sites $\times$ 10 random orderings), we averaged the improvements in localization accuracy conferred by continual learning models at each step within the 29-hospital training sequence (all sites except the held-out site) compared to fine-tuning. Performance improvements increased steadily with additional hospitals across the full training sequence (Spearman $\rho = 0.74$, $P < 0.001$), with no evidence of a plateau (Extended Data Fig. 5). This indicates that continual learning continues to benefit from each additional institutional contributor, suggesting the approach scales favourably with broader participation rather than saturating at a small number of sites.

Lastly, given that additional hospitals generally improved performance, we sought to determine whether the quality of scans provided by individual sites influences the benefit of their inclusion, and whether hospitals contributing lower-quality data should be incorporated into the training sequence at all. One hospital (31st NICU site) contributed radiographs with notably lower resolution than other sites and was accordingly not included within the main experiments. We compared continual learning models trained with and without data from this hospital. Models including this hospital in the training sequence exhibited higher combined localization error (0.75 cm vs 0.67 cm) and achieved positive improvements in precision@15 and recall@15 at only 20 of 30 sites compared to fine-tuning, versus 27 of 30 sites for models excluding these data (Extended Data Fig. 6). These findings suggest that data quality heterogeneity could constrain continual learning benefits, and that quality thresholds for contributing sites may be warranted in future implementations.

Discussion

We present a proof-of-concept framework for repurposing adult radiology AI models to neonatal patients through multihospital collaboration among 30 hospitals across 20 countries and territories, representing one of the most geographically diverse imaging cohorts assembled for neonatal AI development, in a field where most models are developed and validated on single institutions and large-scale multisite datasets remain scarce^{3,33}. Our findings suggest that this approach represents a promising and practical paradigm for model adaptation that does not require hospitals to gather large datasets, share patient data, or enlist specialized AI teams, consistent with previous work on continual learning for medical image interpretation³⁴. Beyond demonstrating

that continual learning improves neonatal ETT assessment, this study systematically characterizes the operating conditions of sequential multisite model adaptation, including the role of adult pretraining, the relationship between the number of contributing sites and performance gains, the impact of data quality heterogeneity, and the retention of cumulative knowledge across the training sequence, providing an evidence base for deploying this paradigm across paediatric imaging tasks more broadly.

A central challenge in multi-institutional AI development is enabling model improvement across sites without compromising patient privacy or requiring prohibitive data governance infrastructure. Other approaches such as centralized data pooling and federated learning each carry substantial operational requirements. Centralized data sharing requires individual data transfer agreements, institutional review board approvals across jurisdictions and compliance with heterogeneous regulatory frameworks, creating practical barriers^{26,35}. Federated learning preserves data locality, but introduces its own complexity: participating sites must maintain synchronized compute infrastructure, sustain communication across training rounds and coordinate heterogeneous local environments, all of which present substantial hurdles to real-world deployment^{26,27}. Our continual learning framework offers a simpler alternative. Although we performed all experiments centrally in this study for practicality, this continual learning framework is designed to allow model training and evaluation locally at each hospital without data exchange. In future implementations, we envision that each hospital could receive a self-contained training container and preprocessing script. Clinical involvement would be limited to curating and annotating a modest dataset, a task that does not require specialized AI expertise. Local IT departments could execute the training pipeline by directing the container to a standardized data directory, verify model performance against pre-set acceptance criteria, and submit the updated model weights and evaluation results to a shared ledger with version control. No dedicated AI team would be required at any individual site. Importantly, ensuring that sequential updates do not degrade performance at previously seen hospitals requires ongoing monitoring. The backward evaluation methodology introduced in this study could serve as the basis for this: after each sequential update, performance on all previously seen hospitals could be assessed against pre-set acceptance criteria, with the shared ledger enabling rollback to previous model versions if regressions are detected.

Our study addresses a critical gap in the paediatric AI literature. Despite the substantial clinical burden of neonatal critical care, paediatric patients are systematically underrepresented in medical AI research³. Adult-trained models consistently demonstrate reduced performance in children, with the greatest deficits in the youngest patients, particularly those under 2 years of age⁹. Furthermore, of the paediatric AI models that have been proposed, the majority rely on single-centre retrospective data; only a small fraction report external validation, and many experience substantial performance drops in cross-institutional settings^{8,33,36}. This lack of rigorous multisite evaluation represents a fundamental barrier to clinical translation, as models that appear accurate on local data may fail when deployed across diverse clinical environments with different imaging equipment, patient demographics and acquisition protocols. Our study demonstrates a proof-of-concept approach validated across 30 NICUs in 20 countries and territories, representing a scale and diversity of evaluation that, to our knowledge, is unprecedented in the paediatric AI literature. Importantly, our underlying approach, using small amounts of locally curated data to sequentially adapt an existing adult model without data sharing, is broadly applicable to any paediatric imaging task for which adult AI models exist but paediatric solutions do not, potentially opening pathways for paediatric AI development that were previously constrained by data scarcity and privacy barriers.

We applied our framework to ETT placement assessment, a task with extensive adult-focused AI models^{22,37–39} yet limited paediatric

solutions. While existing paediatric studies have used clinical and demographic variables to predict ETT size or inappropriate ETT depth^{12,13}, no previous work has performed radiograph-based ETT screening in neonatal populations validated across diverse international settings. We found that continual learning models significantly outperformed fine-tuned and direct prediction models for ETT and carina localization across almost every test site and were more consistently accurate, indicating that continual learning can effectively repurpose a model trained on adult data to neonates for this task. The localization accuracy achieved by continual learning models in our neonatal cohort is comparable to the original model on adult ICU radiographs⁴⁰, despite the substantial anatomical differences between populations. Our backward evaluation confirmed that these improvements did not come at the cost of degraded performance on previously seen hospitals, and instead, reflected a cumulative knowledge built through the continual learning process. In our precision/recall analysis, we found that continual learning models exhibited improved precision and recall and lower false positive rates for isolating low ETT placements compared with baseline methods across the majority of test sites, suggesting that the increase in ETT and carina localization ability conferred by continual learning could translate into clinical predictive benefits. Prospective validation will be essential to confirm these benefits in practice; we envision this to involve integrating the continual learning model into NICU clinical workflows at a subset of participating sites to measure time-to-confirmation for abnormal ETT placements, clinician agreement with model-flagged cases, and downstream impact on re-intubation rates. Such studies could determine whether these models can assist clinicians in prioritizing the most concerning scans for expedited radiologist confirmation, or even trigger communication alerts from the bedside to a radiologist on service.

Several factors may affect the generalizability and clinical applicability of our findings. Our study design required hospitals to provide class-balanced datasets to ensure adequate representation of abnormal tube placements⁴¹, but the performance of continual learning models could change depending on the true prevalence of abnormally placed ETTs. In addition, variability in image quality across participating sites emerged as a constraint on model training, with lower-resolution scans appearing to hinder performance gains from continual learning. Our precision/recall analysis also faces important constraints. The conversion from pixel measurements to centimetre distances relied on estimated ETT tube widths, which probably contributed to imperfect correlations between ETT–carina distances and clinical placement labels, even for low placements where the distance-label correspondence is conceptually direct. While short ETT–carina distances reliably indicate deep intubation, this metric serves as only an indirect proxy for high placements, which are clinically defined by the position of the ETT tip relative to the level of the first thoracic vertebra (estimate of thoracic inlet)⁴². As confirmed by our correlation analyses, the relationship between ETT–carina distance and high placement labels was substantially weaker than for low placements. Future models incorporating localization of additional anatomical landmarks, such as the first thoracic vertebra, could enable more direct assessment of high ETT placements, and the continual learning framework demonstrated here would be directly applicable to such models.

The path towards clinical implementation presents additional challenges. While this study uses retrospective data to establish continual learning as a proof of concept for multi-institutional collaborative model adaptation, prospective trials testing continual learning models integrated directly into clinical workflows will be essential to robustly evaluate their real-world efficacy and potential impact on patient outcomes⁴³. In addition, while our study simulated decentralized training by restricting information transfer to model weights alone, future work should validate this framework under true decentralized execution, where differences in local computing environments and capabilities may introduce additional variability. Furthermore, the evolving nature

of continual learning models may complicate regulatory oversight, as each participating hospital may face unique governance requirements, including institutional review board oversight for using clinical data in model training and internal approval processes for adopting clinical decision-support tools³⁵, that need reconciliation with FDA approval processes. Finally, our study encompassed a diverse international population, and while we did not observe significant pairwise differences in performance stratified across demographic groups, further demographic discrepancies require additional investigation. Relatedly, although model localization accuracy did not significantly differ across gestational age groups in pairwise comparisons, the clinical margin for acceptable ETT positioning is substantially narrower in extremely premature neonates^{44,45}; prospective validation in this population will be essential to determine whether the observed accuracy is clinically sufficient.

Ultimately, this study provides a proof of concept for using continual learning to repurpose adult AI models for neonatal radiograph interpretation, validated across one of the most geographically diverse multi-institutional paediatric imaging cohorts assembled so far. To our knowledge, no previous study has proposed and evaluated a privacy-preserving model adaptation framework across this number of international paediatric sites, or demonstrated head-to-head performance comparisons on neonatal data from such diverse clinical sources. Our results suggest that continual learning offers a practical, data-efficient alternative to centralized training and federated learning for multi-institutional adult to paediatric model adaptations. These findings support further investigation into continual learning as a paradigm to improve the real-world performance of paediatric AI algorithms across different point-of-care settings.

Methods

Data source and patient population

This retrospective study utilized chest radiographs from NICU patients obtained from a global multi-institutional collaboration, approved by the Harvard Longwood Campus Institutional Review Board (HLC IRB) at Harvard Medical School (IRB22-0400), to develop a privacy-focused and data-efficient approach for adapting adult radiograph interpretation models to a paediatric setting. Methodological reporting followed the CLAIM 2024 checklist⁴⁶, attached as a supplement. Each participating institution ensured that data sharing complied with an appropriate consent or policy and obtained institutional review board approval for the data collection and sharing protocol. Patients or the public were not involved in the design, conduct, reporting or dissemination of this research. Every participating institution collected ~100 studies with chest radiographs from NICU patients. A formal sample size calculation was not performed, as no established power analysis frameworks exist for sequential multisite model adaptation studies where performance depends on numerous interacting factors including the number of contributing sites, intersite data heterogeneity and task difficulty. The target of ~100 radiographs per hospital was determined collaboratively with participating institutions as a pragmatic collection target balancing acquisition feasibility across 30 international sites with sufficient data for local model training and evaluation.

Eligible NICU patients were aged 0–6 months, including premature births, and underwent single bedside chest-, chest–abdomen- or anteroposterior-view radiographs. We restricted samples to those collected in 2021, the year data collection began, to reduce temporal shifts and ensure consistency across the geographical sites. When insufficient cases were available from 2021, studies from 2020 and earlier years were included. All clinical indications for imaging were included. Given the small target sample size per institution, each site aimed to collect a balanced representation of ETT positioning for stable learning and reliable evaluation⁴¹: ~50% of cases with correctly positioned ETTs and 50% with incorrectly positioned ETTs, as determined by radiologist interpretation in the original reports or by collaborating physicians

during data collection. Data from each hospital were split into train and test sets, balanced on the basis of the number of correctly and incorrectly positioned ETTs. Supplementary Fig. 1 contains a flowchart that outlines exclusions and the final train and test sets.

While image acquisition protocol varied per institution, for each case, the following data elements were requested: chest radiograph images, technical imaging parameters (including manufacturer, pixel spacing and imaging modality such as computed radiography or digital radiography), radiology reports, patient sex, clinical indication for radiography, gestational age and patient weight. For this study, only the chest radiograph images and ETT and carina annotations were used for object detection. When submitted data had missing elements, institutions were contacted and asked to provide the missing information; cases with persistently unavailable data elements were retained in the study and handled as described in the ‘Data splitting and preprocessing’ section of the Methods and the Supplementary Methods (for example, estimation of ETT width from gestational age when weight was unavailable). To minimize the influence of data-entry errors, demographic fields submitted in heterogeneous free-text formats (for example, gestational age recorded as completed weeks plus days, or weights in differing units) were normalized to common numerical units where unambiguously interpretable; continuous demographic variables were restricted to biologically plausible ranges for summary statistics (birth weight from 0 to 10,000 g and gestational age from 20 to 44 weeks), and free-text values that could not be parsed numerically (for example, ‘Premature’, ‘no records’) were not included.

Data splitting and preprocessing

Every train split across the hospitals had exactly 50 images to ensure standardization across the continual learning process, whereas the test set contained the rest of the images, which was between 43 and 53 images based on the number of viable samples provided by the institution. To accomplish this while maintaining a consistent label distribution, the data were first stratified by ETT placement classification (low/normal/high). For every hospital, the proportion of normal, low and high ETT placements was calculated out of the initial ~100 samples. Each ETT placement group was allocated training samples proportional to its representation in the full set of samples from the hospital (for example, if 50% of images contained normal placements, then ~25 out of the 50 samples in the train set for the hospital of interest would be normal). To ensure that there were exactly 50 images in each train set, initial numbers of normal, high and low samples were allocated on the basis of these ratios. If this initial allocation was less than 50 images, images would be added from ETT placement groups with available samples, prioritizing groups with more samples. In the rarer cases where the initial allocation is greater than 50 images, images would be removed starting from the groups with the smallest training allocations while maintaining at least one sample per ETT placement category. This approach allowed for the creation of class-balanced train/test splits across every hospital with the exact same number of train scans, as shown in Extended Data Table 1. Samples from the 31st NICU site were not included in the final analysis due to low-resolution radiographs: the native image dimensions were smaller than the 640×640 pixel model input size, requiring upscaling rather than the downscaling applied to all other sites, and the radiographic field of view occupied a substantially smaller proportion of the image area compared with radiographs from other institutions. The impacts of this hospital's inclusion on continual learning are outlined in the ‘Evaluating the extents and limits of continual learning’ subsection below, and more information on this hospital is provided in the Supplementary Methods. In addition, there were 5 patients with two different radiograph samples, meaning that there were 1,499 and 1,461 unique patients included in the train and test sets respectively.

Once split into train/test groups, data across the different sites were pooled and preprocessed using a single script to ensure that all of the data preprocessing and quality control steps were performed

identically across sites and demographic subgroups. Image annotations were processed as bounding boxes adhering to the standardized COCO format⁴⁷. The images were normalized using histogram equalization to standardize contrast variations⁴⁸ across imaging equipment and acquisition parameters from different sites, and were subsequently resized to 640×640 pixels via cubic interpolation to match the input resolution of the model architecture⁴⁹, with zero padding applied to preserve the image's original aspect ratio. Greyscale images were converted to 3-channel RGB format by duplicating the single channel to allow compatibility with the pretrained model used. The bounding boxes were also scaled accordingly using the same scaling factor, with out-of-bounds bounding box coordinates clamped to fit in the new image dimensions. A full implementation of this preprocessing can be found in the released source code.

Standard chest radiographs embed pixel spacing metadata in DICOM headers to enable conversion of pixel distances to physical measurements. However, pixel spacing values were inconsistently reported across our 30 international sites and the ones that were reported did not result in reliably accurate distance measurements. To address this, we used the ETT itself as an internal calibration reference. During annotation, annotators marked the width of the ETT in pixels. Since ETT diameter is known and selected on the basis of patient weight, this pixel width was used to derive a patient-specific pixel-to-centimetre conversion factor. In cases where the weight was not reported, we used the gestational age as an estimate for the weight, and if that was also not reported, we used an estimate of 3.5 mm for the ETT width. Further correlation analysis was conducted to verify that this conversion method is reasonably consistent with the delineation between low and normal/high ETTs. These annotations were scaled in the same manner as the other anatomical landmark annotations, and our process for these conversions is formalized in the Supplementary Methods. All data underwent rigorous de-identification procedures, including manual removal of personal identifiers from radiology reports and automated removal of protected health information from imaging metadata. Images containing identifiable features such as dental structures, partial skull views or jewelry were cropped.

Annotation protocol

Image annotation was performed in two rounds with an additional verification step. Six medical students completed the initial round of annotation, with each assigned a non-overlapping subset of approximately one-sixth of the total samples. A third-year radiology resident and a physician with more than 10 years of clinical experience then reviewed and refined all annotations in a second round, with any disagreements resolved by deferring to this higher level of expertise. Finally, a board-certified radiologist with more than 10 years of experience verified the final labels on a random 20% of samples, identifying no errors. Inter-observer variability was not formally assessed as in the cascading review design, each subsequent annotator reviewed and refined previous annotations rather than annotating independently, whereas intra-observer variability could not be assessed as repeat annotations by the same annotator were not performed. Annotation quality was instead ensured through this hierarchical review process in which each successive round took precedence given increasing clinical expertise.

A commercial digital annotation platform was utilized for delineating bounding boxes on NICU chest radiographs, with images randomly sampled for annotation to minimize systematic bias. Annotation guidelines required the annotation to define the minimum bounding box dimensions that encompassed the ETT and carina with high confidence. This approach balanced annotation precision with anatomical structure capture, ensuring that bounding boxes were sufficiently tight to minimize background inclusion while maintaining complete coverage of the target structures. Annotations were also performed on the depth of ETT placement (low, normal or high), as well as the width of

the ETT. While the thoracic vertebrae were annotated, they were not used in the model development process and instead only served as reference landmarks. The specific protocol provided to annotators is included in the Supplementary Methods.

Continual learning and baselines

In continual learning, an initial model is passed from hospital to hospital sequentially, where each hospital trains the model on its dataset within its own compute environment, without needing to share data, and is only required to share the model weight externally. Our framework employs sequential model updates without explicit forgetting mitigation strategies; instead, we empirically verify that cumulative knowledge is retained through a backward evaluation of model performance across the training sequence (see ‘Model evaluation’ subsection). For each hospital evaluated, we trained a continual learning model on all of the preceding hospitals and then measured performance on an unseen target hospital. Here, ‘target hospital’ refers to the hospital of interest that the model is being evaluated on. In our case, we constructed 30 different continual learning models using each of our 30 hospitals as the unseen target hospital one time, and averaged results across 10 random hospital orderings per continual learning model, as the order of hospitals used for training may affect results. We performed all experiments centrally in our computing environment for this study to simulate decentralized execution. The methods described here, including the standardized preprocessing, fixed training protocol and weight-only transfer, are designed to be directly applicable in a decentralized setting where each hospital trains locally on its own infrastructure.

We implemented two baseline training methods for comparison. We first evaluated the original model, which has already been trained on a diverse set of adult ICU data, directly on the NICU data (direct prediction). Next, we evaluated against the conventional approach for adapting models, where the model is trained on a small set of NICU data obtained from the target hospital, which we call fine-tuning. A summary of continual learning, fine-tuning and direct prediction is shown in Fig. 1b. Both fine-tuning and direct prediction models were evaluated on the test sets for each hospital, with the fine-tuning models trained using the train set from the target hospital.

Model development

For this study, we adapted the CarinaNet model⁴⁰, a RetinaNet-based⁴⁹ object detection model with a ResNet-50 backbone and feature pyramid network (FPN), pretrained on the COCO dataset⁴⁷ and then further adapted using 1,890 adult ICU radiographs. The model detects two classes, ETT tip and carina, and outputs bounding boxes from which centre-point coordinates are extracted as the predicted landmark locations. We followed the training process detailed in the original CarinaNet documentation⁴⁰. The training loss combines focal loss for classification with smooth L1 loss for bounding box regression, with asymmetric weighting that prioritizes localization of the carina (1.2×)—the anatomical reference for assessing tube placement—over the ETT tip (0.8×). We used the AdamW optimizer with a OneCycleLR scheduler (max learning rate = 2.95×10^{-4}), gradient clipping (max norm=0.1) and batch size of 16. Batch normalization layers were frozen during training. Data augmentation consisted of random rotation ($\pm 10^\circ$) and random vertical cropping (0–25%) with corresponding bounding box coordinate adjustment. Given the limited training neonatal data available at each site, we performed preliminary cross-validation using the 50 train samples to determine the optimal length of training. This process involved evaluating the model at every training epoch on small subsections of the training dataset, and this process was performed for each hospital within the continual learning sequence. The training processes for continual learning and fine-tuning models at an individual hospital were identical. More details on the training and model prediction process can be found in the Supplementary Methods.

Model evaluation

We first compared the ability for models to localize the carina and the ETT tip, reporting errors in median centimetres due to the propensity for the benchmark methods to output high error predictions. Localization errors were summarized as median values per hospital that were averaged across simulations for continual learning and fine-tuning. Combined ETT tip and carina localization error was computed as the sum of the two individual landmark errors for each radiograph. We conducted further testing to evaluate whether continual learning reduces the prevalence of extreme errors, which we define as anatomical localization errors exceeding 5 cm. This threshold was informed by previous radiographic measurements showing that total tracheal length in neonates (22–42 weeks gestation and <1 week of age) ranges from –2.2–6.4 cm, depending on gestational age and birthweight⁴⁴. Although this reference population is not identical to ours (neonates aged 0–6 months, including premature infants), it provides relevant anatomical context: a 5 cm error would approximate or exceed the entire tracheal length in most neonates. The extreme-error threshold of 5 cm was applied to the combined ETT and carina error value. The improvement conferred by continual learning in reducing such errors is visualized in the results in comparison to both fine-tuning and direct prediction.

To directly evaluate whether improvements in ETT and carina localization from continual learning indicated clinically relevant benefits, we measured the ability of continual learning models to triage and flag abnormal ETT placements for further radiologist or clinician review. ETT–carina distances were computed on the basis of the pixel distance between the predicted ETT and carina locations, and these pixel distances were converted to centimetre measures using the pixel width of the ETT as a scale. We conducted correlation analyses using Kendall’s tau⁵⁰ correlations to determine how closely ETT–carina distances corresponded with clinician-assigned placement labels, separately for low and high ETT placements, visualized in Supplementary Figs. 2 and 3. We then evaluated the precision@*k*, recall@*k* and the false positive rate (FPR@*k*) at each hospital, measuring how many of the *k* most concerning radiographs ranked by each model based on predicted ETT–carina distances contained abnormally low or high intubations. For low placements, radiographs were ranked by ascending predicted distance, while for high ETT placements, radiographs were ranked by descending predicted distance. We evaluated the average improvement in these metrics for every hospital, conferred by continual learning across the different random hospital orderings. While continual learning models were compared to the fine-tuning and direct prediction baselines throughout these evaluations, we focused primarily on visualizing the improvements over fine-tuning, as direct prediction performed considerably less consistently. Comparisons among all three are shown in Supplementary Tables 2 and 5–7 as well as Extended Data Figs. 1 and 2. Given the weaker concordance between ETT–carina distance and high placement labels, we prioritized detecting low intubations, with the results for triaging high placements shown in Supplementary Table 7. Lastly, to provide further interpretability, we additionally reported the mean predicted and ground-truth ETT–carina distances for the true positive and false positive cases within the top 15 flagged radiographs, pooled across the 10 continual learning simulations for each hospital to obtain robust estimates (Supplementary Table 8).

To assess whether sequential training resulted in catastrophic forgetting³⁰, which in this context we define as the degradation of performance on hospitals previously seen in the continual learning sequence after subsequent updates, we conducted a backward evaluation of model performance across the training sequence. For a set of 100 random hospital orderings, we evaluated the continual learning model on each hospital’s held-out test set at two timepoints: immediately after training on that hospital (the intermediate model) and after completing the full training sequence (the final model). We computed

the difference in combined ETT and carina localization error between the final and intermediate models for each hospital at each sequence position, averaged across the 100 random orderings. We assessed the relationship between sequence position and performance change using Spearman rank correlation.

Evaluating the extents and limits of continual learning

Adapting AI models trained on adult patients to different NICU sites is difficult for two key reasons. First, models trained on adult data must perform the same task on neonates, a patient population with substantially different anatomical features. Second, data collection protocols can vary between different hospitals, especially internationally, and models must adapt to specific aspects of a hospital's data distribution, such as its scanner, exposure and so on. With this in mind, we conducted several additional experiments to gain better insights on why continual learning leads to improved model performance.

First, we evaluated whether the continual learning models performed better with or without initial training on an extensive set of adult ICU data for ETT and carina annotation, with the assumption that heuristics for the task learned in the ICU could transfer to the NICU. To do so, we performed continual learning with the initial RetinaNet model (not trained on any chest X-rays) and compared its performance to our standard continual learning method with the CarinaNet model (trained on adult chest X-rays). Next, we sought to determine whether there was a point at which adding further hospitals to the continual learning sequence did not lead to meaningful gains in performance versus fine-tuning. At each position within the continual learning sequence, we measured the performance improvement of continual learning models over fine-tuning. These performance improvements were averaged across 300 held-out-site and ordering combinations (30 sites $\times$ 10 random orderings) to evaluate the effect of different numbers of collaborating hospitals within continual learning. Lastly, we tested the continual learning protocol both with and without the low-resolution scans from one of the contributing hospitals that were excluded from the primary analysis to determine whether the quality of scans from a single hospital within the continual learning sequence can have meaningful impacts on the overall model performance.

Statistical analysis

Non-parametric data were summarized as median (with interquartile range [IQR]), and parametric data were summarized as mean $\pm$ s.d. Two-sample comparisons were assessed with the two-sample *t*-test, and one-sample differences with the one-sample *t*-test. For ≥ 3 independent groups, overall differences were evaluated using analysis of variance (ANOVA); when significant ($P < 0.05$), pairwise comparisons were performed using *t*-tests with Holm–Sidak³¹ adjustment. We computed 95% confidence intervals for relevant data ranges across the three methods, reported in parentheses next to outcome values. The confidence intervals reflect different sources of variability for each method: for continual learning, intervals were derived from the distribution of results across random hospital orderings, capturing variability due to training sequence order; for fine-tuning, intervals were derived from independent runs with different random seeds, capturing variability due to training stochasticity; and for direct prediction, which produces a single deterministic output, we computed bootstrap³² confidence intervals ($B = 10,000$) by resampling test images with replacement and recomputing the metric for each resample, with the 2.5th and 97.5th percentiles defining the interval, capturing test-sample uncertainty. Per-hospital comparisons between methods were treated as independent site-level tests and reported without multiple-comparison correction. Statistical analysis was performed using Python v.3.9.19.

Reporting summary

Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

Data availability

The chest radiograph images that support the findings of this study are currently not publicly available because the data use agreements governing this multi-institutional international collaboration prohibit the transfer or redistribution of patient radiographs and patient privacy protections preclude their release. Aggregate summary statistics and demographic characteristics necessary to interpret the findings are provided in Extended Data Tables 1 and 2 and Supplementary Table 1. The trained model weights, which enable application of the continual learning framework to new datasets, are publicly available via GitHub at <https://github.com/rajpurkarlab/MAIDA-NICU-continual-learning> (ref. 53). Source data are provided with this paper.

Code availability

Source code for the analyses in this study is available via GitHub at <https://github.com/rajpurkarlab/MAIDA-NICU-continual-learning> (ref. 53).

References

1. Brewster, R. C., Nagy, M., Wunnava, S. & Bourgeois, F. T. US FDA approval of pediatric artificial intelligence and machine learning-enabled medical devices. *JAMA Pediatr.* **179**, 212–214 (2025).
2. Muralidharan, V. et al. A scoping review of reporting gaps in FDA-approved AI medical devices. *NPJ Digit. Med.* **7**, 273 (2024).
3. Sammer, M. B. et al. Use of artificial intelligence in radiology: impact on pediatric patients, a white paper from the ACR pediatric AI workgroup. *J. Am. Coll. Radiol.* **20**, 730–737 (2023).
4. Shin, H. J., Son, N.-H., Kim, M. J. & Kim, E.-K. Diagnostic performance of artificial intelligence approved for adults for the interpretation of pediatric chest radiographs. *Sci. Rep.* **12**, 10215 (2022).
5. Hua, S. B. Z. et al. Underrepresentation of children in public medical imaging datasets. *Nat. Health* **1**, 888–897 (2026).
6. Ramgopal, S. et al. Artificial intelligence-based clinical decision support in pediatrics. *Pediatr. Res.* **93**, 334–341 (2023).
7. Ciet, P. et al. The unintended consequences of artificial intelligence in paediatric radiology. *Pediatr. Radiol.* **54**, 585–593 (2024).
8. Wang, Q. et al. Applications of artificial intelligence in early childhood health management: a systematic review from fetal to pediatric periods. *Front. Pediatr.* **13**, 1613150 (2025).
9. Laborie, L. B., Lee, J., Antram, E., Lein, R. K. & Shelmerdine, S. C. Performance of adult-trained artificial intelligence models in paediatric imaging—a scoping review. *Eur. Radiol.* **36**, 5916–5939 (2026).
10. Hwang, E. J., Goo, J. M. & Park, C. M. AI applications for thoracic imaging: considerations for best practice. *Radiology* **314**, e240650 (2025).
11. GE Medical Systems. 510(k) Premarket Notification K211161: Critical Care Suite with Endotracheal Tube Positioning AI https://www.accessdata.fda.gov/cdrh_docs/pdf21/K211161.pdf (US FDA, 2021).
12. Zhou, M. et al. Prediction of endotracheal tube size in pediatric patients: development and validation of machine learning models. *Front. Pediatr.* **10**, 970646 (2022).
13. Shim, J.-G. et al. Predicting the risk of inappropriate depth of endotracheal intubation in pediatric patients using machine learning approaches. *Sci. Rep.* **13**, 5156 (2023).
14. Foglia, E. E. et al. Factors associated with adverse events during tracheal intubation in the NICU. *Neonatology* **108**, 23–29 (2015).
15. Hatch, L. D. et al. Endotracheal intubation in neonates: a prospective study of adverse safety events in 162 infants. *J. Pediatr.* **168**, 62–66.e6 (2016).
16. Sauer, C. W. et al. Intubation attempts increase the risk for severe intraventricular hemorrhage in preterm infants—a retrospective cohort study. *J. Pediatr.* **177**, 108–113 (2016).

17. Miller, K. A., Kimia, A., Monuteaux, M. C. & Nagler, J. Factors associated with misplaced endotracheal tubes during intubation in pediatric patients. *J. Emerg. Med.* **51**, 9–18 (2016).
18. Harris, E. A., Arheart, K. L. & Penning, D. H. Endotracheal tube malposition within the pediatric population: a common event despite clinical evidence of correct placement. *Can. J. Anaesth.* **55**, 685–690 (2008).
19. Rohana, J., Boo, N. Y., Yong, S. C. & Ong, L. C. Turn-around-time of radiographs in a neonatal intensive care unit. *Med. J. Malaysia* **60**, 338–344 (2005).
20. Fink, A. Z. et al. Discrepancies in radiograph interpretation between pediatric radiologists and pediatric intensivists in the pediatric or neonatal intensive care unit. *J. Am. Coll. Radiol.* **15**, 1580–1586 (2018).
21. Novak, A. et al. Evaluation of the impact of artificial intelligence-assisted image interpretation on the diagnostic performance of clinicians in identifying endotracheal tube position on plain chest X-ray: a multi-case multi-reader study. *Crit. Care* **29**, 330 (2025).
22. An, J. Y. et al. Artificial intelligence for assessment of endotracheal tube position on chest radiographs: validation in patients from two institutions. *Am. J. Roentgenol.* **222**, e2329769 (2024).
23. van de Ven, G. M., Tuytelaars, T. & Tolias, A. S. Three types of incremental learning. *Nat. Mach. Intell.* **4**, 1185–1197 (2022).
24. Wang, L., Zhang, X., Su, H. & Zhu, J. A comprehensive survey of continual learning: theory, method and application. *IEEE Trans. Pattern Anal. Mach. Intell.* **46**, 5362–5383 (2024).
25. Chen, Z. & Liu, B. *Lifelong Machine Learning* (Morgan & Claypool Publishers, 2018).
26. Rieke, N. et al. The future of digital health with federated learning. *NPJ Digit. Med.* **3**, 119 (2020).
27. Sheller, M. J. et al. Federated learning in medicine: facilitating multi-institutional collaborations without sharing patient data. *Sci. Rep.* **10**, 12598 (2020).
28. Darzidehkalani, E., Ghasemi-Rad, M. & van Ooijen, P. M. A. Federated learning in medical imaging: part II: methods, challenges, and considerations. *J. Am. Coll. Radiol.* **19**, 975–982 (2022).
29. Chang, K. et al. Distributed deep learning networks among institutions for medical imaging. *J. Am. Med. Inform. Assoc.* **25**, 945–954 (2018).
30. Kirkpatrick, J. et al. Overcoming catastrophic forgetting in neural networks. *Proc. Natl Acad. Sci. USA* **114**, 3521–3526 (2017).
31. Pinheiro, J. M., Munshi, U. K. & Chowdhry, R. Strategies to improve neonatal intubation safety by preventing endobronchial placement of the tracheal tube—literature review and experience at a tertiary center. *Children* **10**, 361 (2023).
32. Volsko, T. A., McNinch, N. L., Prough, D. S. & Bigham, M. T. Adherence to endotracheal tube depth guidelines and incidence of malposition in infants and children. *Respir. Care* **63**, 1111–1117 (2018).
33. Padash, S., Mohebbian, M. R., Adams, S. J., Henderson, R. D. & Babyn, P. Pediatric chest radiograph interpretation: how far has artificial intelligence come? A systematic literature review. *Pediatr. Radiol.* **52**, 1568–1580 (2022).
34. Bruno, P., Quarta, A. & Calimeri, F. Continual learning in medicine: a systematic literature review. *Neural Process. Lett.* **57**, 2 (2025).
35. Price, W. N. & Cohen, I. G. Privacy in the age of medical big data. *Nat. Med.* **25**, 37–43 (2019).
36. Elahmedi, M., Sawhney, R., Guadagno, E., Botelho, F. & Poenaru, D. The state of artificial intelligence in pediatric surgery: a systematic review. *J. Pediatr. Surg.* **59**, 774–782 (2024).
37. Lakhani, P., Flanders, A. & Gorniak, R. Endotracheal tube position assessment on chest radiographs using deep learning. *Radiol. Artif. Intell.* **3**, e200026 (2021).
38. Brown, M. S. et al. Automated endotracheal tube placement check using semantically embedded deep neural networks. *Acad. Radiol.* **30**, 412–420 (2023).
39. Lederman, D. An artificial intelligence system for endotracheal intubation confirmation. In *International Conference on Engineering Applications of Neural Networks* (eds Iliadis, L. et al.) 133–140 (Springer, 2021).
40. Oliver, M. et al. Image augmentation and automated measurement of endotracheal-tube-to-carina distance on chest radiographs in intensive care unit using a deep learning model with external validation. *Crit. Care* **27**, 40 (2023).
41. Mazurowski, M. A. et al. Training neural network classifiers for medical decision making: the effects of imbalanced datasets on classification performance. *Neural Netw.* **21**, 427–436 (2008).
42. Blayney, M. P. & Logan, D. R. First thoracic vertebral body as reference for endotracheal tube placement. *Arch. Dis. Child. Fetal Neonatal Ed.* **71**, F32–F35 (1994).
43. Kelly, C. J., Karthikesalingam, A., Suleyman, M., Corrado, G. & King, D. Key challenges for delivering clinical impact with artificial intelligence. *BMC Med.* **17**, 195 (2019).
44. Cerone, J. B. & Pinheiro, J. M. Tracheal length measurement in intubated neonates to guide the design and use of endotracheal tube glottic depth markings. *Children* **9**, 169 (2022).
45. Sakhuja, P., Finelli, M., Hawes, J. & Whyte, H. Is it time to review guidelines for ETT positioning in the NICU? SCEPTIC—survey of challenges encountered in placement of endotracheal tubes in Canadian NICUs. *Int. J. Pediatr.* **2016**, 7283179 (2016).
46. Tejani, A. S. et al. Checklist for artificial intelligence in medical imaging (CLAIM): 2024 update. *Radiol. Artif. Intell.* **6**, e240300 (2024).
47. Lin, T.-Y. et al. Microsoft COCO: common objects in context. In *European Conference on Computer Vision* (eds Fleet, D. et al.) 740–755 (Springer, 2014).
48. van der Walt, S. et al. scikit-image: image processing in Python. *PeerJ* **2**, e453 (2014).
49. Lin, T.-Y., Goyal, P., Girshick, R., He, K. & Dollár, P. Focal loss for dense object detection. In *Proc. IEEE International Conference on Computer Vision* 2999–3007 (IEEE, 2017).
50. Kendall, M. G. A new measure of rank correlation. *Biometrika* **30**, 81–93 (1938).
51. Holm, S. A. A simple sequentially rejective multiple test procedure. *Scand. J. Stat.* **6**, 65–70 (1979).
52. Efron, B. & Tibshirani, R. J. *An Introduction to the Bootstrap* (Chapman & Hall, 1993).
53. Rao, V. M. MAIDA-NICU-continual-learning (v1.0.0). *Zenodo* <https://doi.org/10.5281/zenodo.22945721> (2026).

Acknowledgements

We thank W. Erselius for managing the partnerships, and C. Perry and J. Sullivan for their support. We thank Microsoft Research, and K. Takeda in particular, for supporting our work. We thank the Department of Neonatology at Dana Dwek Children's Hospital, Tel Aviv Sourasky Medical Center, Tel Aviv, Israel, for their support in this study.

Author contributions

V.M.R. led the study, with E. Chen and A.S. as co-leads. V.M.R. contributed to conceptualization, methodology, software, validation, formal analysis, investigation, data curation, writing of the original manuscript draft, visualization and project administration. E. Chen contributed to conceptualization, methodology, software, validation, investigation, data curation, writing of the original manuscript draft, review and editing, and project administration. A.S. contributed to conceptualization, writing of the original draft, review and editing, and project administration. O.B. contributed to conceptualization, software,

investigation, writing of the original draft, and review and editing. H.M. contributed to conceptualization and data curation. L.B. contributed to conceptualization, methodology, validation, investigation and data curation. S.E.K. contributed to conceptualization, methodology, software, validation, investigation and data curation. E.Z. contributed to conceptualization, methodology, validation, investigation and data curation. S. Shelmerdine contributed to conceptualization, methodology, investigation and data curation. J.P., K.R.K., S. Suryadevara, S.S.D.J., J.L. and R.A. contributed to data curation, and paper review and editing. M. Scala, J.F.N., N.S., S. Tschauner, G.M.S., M.J.V.V., J.S.O.V., O.J.R., O.J.L.U., S. Afat, T.K., S.T.A., D.T., K.L.P., J.K., N.H.A., R.M.I., R.S., A.E., I.A., M.B., Y.M., D.U., A.S.A.-G., S.A.Z., M.A.S.M., W.I.B.W.M., N.N., R.K., K.A.-K., Z.I., A.A.S., A.K., T.P., N.P., D.S.W.L., J.T.P.D.H., P.N., M.L.C., M.B.-Z., A.A.-O., A.N., Y.-C.L., W.-T.L., W.-H.F., C.L., A.M., S. Sriswasdi, P.K., S. Tanmuk, D.A., S.C., S. Alan, U.A.T., D.A.M., E. Chung, M. Sadler, M.P.M. and R.H.J. contributed to investigation and resources. P.R. supervised the study and contributed to conceptualization, methodology, validation, resources, writing of the original draft, review and editing, supervision, project administration and funding acquisition. All authors reviewed and edited, and approved the final manuscript.

Funding

This work was supported by Biswas Family Foundation's Transformative Computational Biology Grant in Collaboration with the Milken Institute. L.B. was supported by Junior Orsi and Barco. The funders had no role in study design, data collection and analysis, decision to publish or preparation of the manuscript.

Competing interests

D.T. received honoraria from Bayer, GE, Roche, AstraZeneca, MSD, Gilead, and Philips, and holds shares in StratifAI GmbH, Germany, and in Synagen GmbH, Germany. The other authors declare no competing interests.

Additional information

Extended data is available for this paper at <https://doi.org/10.1038/s41551-026-01812-6>.

Supplementary information The online version contains supplementary material available at <https://doi.org/10.1038/s41551-026-01812-6>.

Correspondence and requests for materials should be addressed to Pranav Rajpurkar.

Peer review information *Nature Biomedical Engineering* thanks Kristyn Beam, Amit Gupta and Curtise Ng for their contribution to the peer review of this work. Peer reviewer reports are available.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2026

Vishwanatha M. Rao^{1,2,3,60}, Emma Chen^{1,4,60}, Agustina Saenz^{1,5,60}, Oishi Banerjee^{1,4}, Henrik Marklund⁶, Lucas Bijmens^{1,7}, Sung Eun Kim¹, Evan Zucker⁸, Susan Shelmerdine⁹, Jeremy Palacio¹⁰, Kent Ryan Kleinschmidt¹⁰, Sathvik Suryadevara¹⁰, Sri Sai Dinesh Jaliparthi¹⁰, Jiaxi Lu¹⁰, Rithvik Akula¹⁰, Mario Scala^{11,12}, Julian F. Niedermair^{11,12}, Nikolaus Stranger¹³, Sebastian Tschauner¹³, Georg M. Schmölzer¹⁴, Maria Jose Veloza Vega¹⁵, Juan Sebastián Osorio Valencia^{15,16}, Oscar Jaramillo Robledo¹⁷, Oscar Julian López Uribe¹⁸, Saif Afat¹⁹, Thomas Küstner²⁰, Soroosh Tayebi Arasteh^{21,22}, Daniel Truhn²³, Kelsey L. Pomykala²⁴, Jens Kleesiek²⁴, Naela Himayati Afifah²⁵, Rozan M. Irfan²⁵, Reza Shahriarirad²⁶, Amirhossein Erfani²⁶, Israel Amirav^{14,27,28}, Moria Be'er^{27,28}, Yasuhito Mitsuyama²⁹, Daiju Ueda³⁰, Alain S. Abi-Ghanem³¹, Safaa Al Zakleet³¹, Mossad Abdelhak Shaban Mohamed³², Wan Irfan Bin W Mustapha³³, Naila Nizami³⁴, Rafeah Khan³⁵, Khalid Al-Kharazi³⁶, Zafar Iqbal³⁷, Abdullah Al Shahrani³⁸, Amna Kashgari³⁹, Tijana Prodanovic⁴⁰, Nikola Prodanovic⁴¹, Desmond Shi Wei Lim⁴², James Thomas Patrick Decourcy Hallinan⁴², Phumza Nongena⁴³, Melissa Louise Cooke⁴³, Montserrat Bret-Zurita⁴⁴, Almudena Alonso-Ojembarrena⁴⁵, Alexandra Nemes⁴⁶, Yung-Chieh Lin⁴⁷, Wei-Ting Lin⁴⁷, Wen-Hui Fang⁴⁸, Chin Lin⁴⁹, Atchara Mahayosnond⁵⁰, Sira Sriswasdi⁵¹, Piyapong Khumrin^{52,53}, Suphakris Tanmuk⁵⁴, Didem Arman⁵⁵, Serdar Cömert⁵⁵, Serdar Alan⁵⁶, Umit Ayse Tandircioglu⁵⁶, David Andrew Mong⁵⁷, Esther Chung⁵⁸, Michael Sadler⁵⁸, Morgan P. McBee⁵⁹, Richard H. Jones⁵⁹ & Pranav Rajpurkar¹✉

¹Department of Biomedical Informatics, Harvard Medical School, Boston, MA, USA. ²Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA, USA. ³Department of Bioengineering, School of Engineering and Applied Science, University of Pennsylvania, Philadelphia, PA, USA.

⁴School of Engineering and Applied Sciences, Harvard University, Cambridge, MA, USA. ⁵Department of Medicine, Brigham and Women's Hospital, Boston, MA, USA. ⁶Department of Computer Science, Stanford University, Stanford, CA, USA. ⁷Department of Imaging and Pathology, KU Leuven, Leuven, Belgium. ⁸Department of Radiology/Pediatric Radiology, Stanford University, Palo Alto, CA, USA. ⁹Department of Clinical Radiology, Great Ormond Street Hospital for Children, London, UK. ¹⁰Saint Louis University School of Medicine, Saint Louis University, Saint Louis, MO, USA. ¹¹Central Radiology Institute, Johannes Kepler University Linz, Kepler University Hospital GmbH, Linz, Austria. ¹²Division of Virtual Morphology, Institute of Anatomy and Cell Biology, Johannes Kepler University, Linz, Austria. ¹³Division of Pediatric Radiology, Department of Radiology, Medical University of Graz, Graz, Austria.

¹⁴Department of Pediatrics, Faculty of Medicine and Dentistry, University of Alberta, Edmonton, Alberta, Canada. ¹⁵Artificial Intelligence Committee, Colombian Association of Radiology, Medellín, Colombia. ¹⁶Medical Imaging and AI in Health SURA, Bioscience Center, Ayudas Diagnósticas SURA, Medellín, Colombia. ¹⁷Education and Research Department, Servicios Especiales de Salud Hospital Universitario de Caldas, Manizales, Colombia. ¹⁸Department of Pediatrics and Neonatology, Servicios Especiales de Salud Hospital Universitario de Caldas, Manizales, Colombia. ¹⁹Department of Diagnostic and Interventional Radiology, Eberhard Karls University, Tuebingen University Hospital, Tuebingen, Germany. ²⁰Medical Image and Data Analysis (MIDAS) Laboratory, University Hospital of Tübingen, Tübingen, Germany. ²¹Department of Diagnostic and Interventional Radiology, University Hospital RWTH Aachen, Aachen, Germany. ²²Department of Urology, Stanford University, Stanford, CA, USA. ²³Lab for AI in Medicine, Department of Diagnostic and Interventional Radiology, University Hospital RWTH Aachen, Aachen, Germany. ²⁴Institute for AI in Medicine (IKIM), University Medicine Essen, Essen, Germany. ²⁵Department of Radiology, RSUP Dr Sardjito, Faculty of Medicine, Universitas Gadjah Mada, Yogyakarta, Indonesia. ²⁶Thoracic and Vascular Surgery Research Center, Namazi Hospital, Shiraz University of Medical Sciences, Shiraz, Iran. ²⁷Pediatric Pulmonology Institute, Dana Children's Hospital, Tel Aviv Sourasky University Medical Center (Ichilov), Tel Aviv, Israel. ²⁸Gray Faculty of Medical and Health Sciences, Tel Aviv University, Tel Aviv, Israel. ²⁹Department of Diagnostic and Interventional Radiology, Graduate School of Medicine, Osaka Metropolitan University, Osaka, Japan. ³⁰Department of Artificial Intelligence, Graduate School of Medicine, Osaka Metropolitan University, Osaka, Japan. ³¹Department of Diagnostic Radiology, American University of Beirut, Beirut, Lebanon. ³²Department of Paediatrics, Sultan Ahmad Shah Medical Centre (SASMEC), International Islamic University Malaysia, Kuantan, Malaysia. ³³Department of Radiology, Sultan Ahmad Shah Medical Centre @IIUM, International Islamic University Malaysia, Kuantan, Malaysia. ³⁴Department of Neonatal Pediatrics, The Indus Hospital and Health Network, Karachi, Pakistan. ³⁵Department of Diagnostic Radiology, The Indus Hospital and Health Network, Karachi, Pakistan. ³⁶Department of Neurosurgery, Sidra Medicine, Doha, Qatar. ³⁷Department of Radiology, Sidra Medicine, Doha, Qatar. ³⁸Manager of Diagnostic Imaging Section, King Abdul-Aziz Medical City, Ministry of National Guard Health Affairs, Riyadh, Saudi Arabia. ³⁹Pediatric Radiology Consultant, King Abdullah Specialized Children Hospital, Riyadh, Saudi Arabia. ⁴⁰Department of Pediatrics, Faculty of Medical Sciences, University of Kragujevac, Kragujevac, Serbia. ⁴¹Department of Surgery, Faculty of Medical Sciences, University of Kragujevac, Kragujevac, Serbia. ⁴²Department of Diagnostic Imaging, National University Hospital Singapore, Singapore, Singapore. ⁴³Department of Paediatrics, New Somerset Hospital, University of Cape Town, Cape Town, South Africa. ⁴⁴Department of Radiology, La Paz University Hospital, Madrid, Spain. ⁴⁵Neonatal Intensive Care Unit, Biomedical Research and Innovation Institute of Cádiz (INiBICA), Research Unit, Puerta del Mar University Hospital, Cádiz, Spain. ⁴⁶Department of Neonatology, Life Memorial Hospital, Titu Maiorescu Faculty of Medicine, Bucharest, Romania. ⁴⁷Department of Pediatrics, National Cheng Kung University Hospital, College of Medicine, National Cheng Kung University, Tainan, Taiwan. ⁴⁸Department of Family and Community Medicine, Tri-Service General Hospital, National Defense Medical University, Taipei, Taiwan. ⁴⁹Medical Technology Education Center, School of Medicine, College of Medicine, National Defense Medical University, Taipei, Taiwan. ⁵⁰Department of Radiology, Faculty of Medicine, Chulalongkorn University, Bangkok, Thailand. ⁵¹Center for AI in Medicine, Research Affairs, Faculty of Medicine, Chulalongkorn University, Bangkok, Thailand. ⁵²Department of Family Medicine, Faculty of Medicine, Chiang Mai University, Chiang Mai, Thailand. ⁵³Bioinformatics Core, Luxembourg Centre for Systems Biomedicine (LCSB), University of Luxembourg, Esch-sur-Alzette, Luxembourg. ⁵⁴Ministry of Public Health, Maharaj Nakorn Chiang Mai Hospital, Chiang Mai, Thailand. ⁵⁵Department of Pediatrics, Division of Neonatology, Istanbul Training and Research Hospital, Istanbul, Turkey. ⁵⁶Department of Pediatrics, Division of Neonatology, Kirikkale University School of Medicine, Kirikkale, Turkey. ⁵⁷University of Colorado School of Medicine, Children's Hospital Colorado, Aurora, CO, USA. ⁵⁸Department of Radiology, Newark Beth Israel Medical Center, RWJ Barnabas Health, Newark, NJ, USA. ⁵⁹Department of Radiology and Radiological Science, Medical University of South Carolina, Charleston, SC, USA. ⁶⁰These authors contributed equally: Vishwanatha M. Rao, Emma Chen, Agustina Saenz. ✉ e-mail: pranav_rajpurkar@hms.harvard.edu

Extended Data Table 1 | Overview of the train and test set compositions across the NICU sites

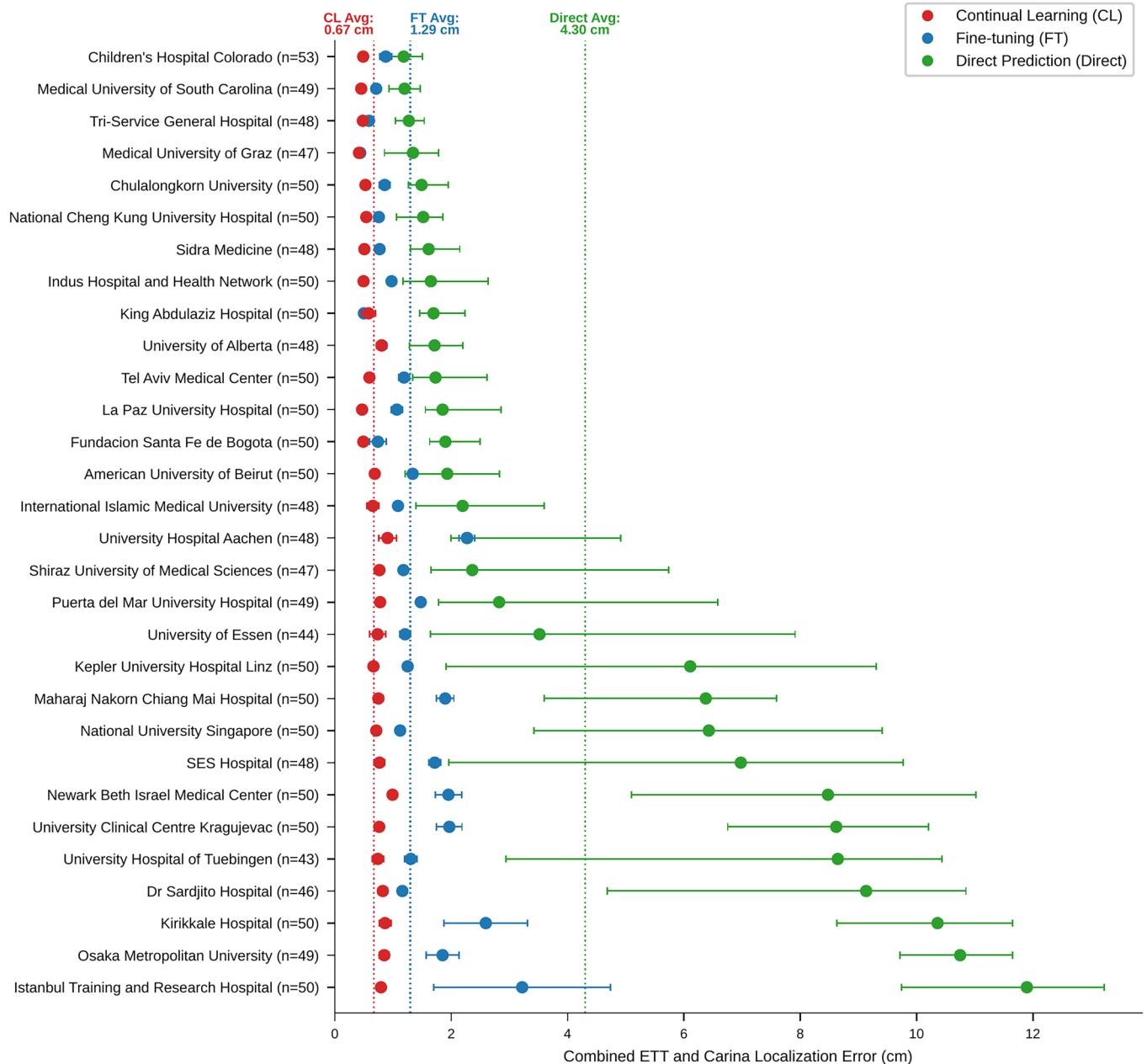
Hospital	Total Images	Training Set				Test Set			
		n	Low n (%)	Normal n (%)	High n (%)	n	Low n (%)	Normal n (%)	High n (%)
American University of Beirut	100	50	25 (50.0)	19 (38.0)	6 (12.0)	50	23 (46.0)	19 (38.0)	8 (16.0)
Children's Hospital Colorado	103	50	32 (64.0)	14 (28.0)	4 (8.0)	53	32 (60.4)	16 (30.2)	5 (9.4)
Chulalongkorn University	100	50	17 (34.0)	28 (56.0)	5 (10.0)	50	18 (36.0)	27 (54.0)	5 (10.0)
Dr Sardjito Hospital	96	50	34 (68.0)	15 (30.0)	1 (2.0)	46	30 (65.2)	14 (30.4)	2 (4.3)
Fundacion Santa Fe de Bogota	100	50	11 (22.0)	28 (56.0)	11 (22.0)	50	11 (22.0)	27 (54.0)	12 (24.0)
Indus Hospital and Health Network	100	50	26 (52.0)	18 (36.0)	6 (12.0)	50	24 (48.0)	19 (38.0)	7 (14.0)
International Islamic Medical University	98	50	21 (42.0)	25 (50.0)	4 (8.0)	48	22 (45.8)	22 (45.8)	4 (8.3)
Istanbul Training and Research Hospital	100	50	15 (30.0)	27 (54.0)	8 (16.0)	50	15 (30.0)	27 (54.0)	8 (16.0)
Kepler University Hospital Linz	100	50	20 (40.0)	24 (48.0)	6 (12.0)	50	20 (40.0)	23 (46.0)	7 (14.0)
King Abdulaziz Hospital	100	50	18 (36.0)	26 (52.0)	6 (12.0)	50	19 (38.0)	25 (50.0)	6 (12.0)
Kirikkale Hospital	100	50	24 (48.0)	20 (40.0)	6 (12.0)	50	23 (46.0)	20 (40.0)	7 (14.0)
La Paz University Hospital	100	50	15 (30.0)	25 (50.0)	10 (20.0)	50	16 (32.0)	24 (48.0)	10 (20.0)
Maharaj Nakorn Chiang Mai Hospital	100	50	16 (32.0)	22 (44.0)	12 (24.0)	50	17 (34.0)	20 (40.0)	13 (26.0)
Medical University of Graz	97	50	30 (60.0)	17 (34.0)	3 (6.0)	47	28 (59.6)	16 (34.0)	3 (6.4)
Medical University of South Carolina	99	50	32 (64.0)	18 (36.0)	0 (0.0)	49	30 (61.2)	19 (38.8)	0 (0.0)
National Cheng Kung University Hospital	100	50	25 (50.0)	23 (46.0)	2 (4.0)	50	24 (48.0)	23 (46.0)	3 (6.0)
National University Singapore	100	50	23 (46.0)	21 (42.0)	6 (12.0)	50	23 (46.0)	21 (42.0)	6 (12.0)
New Somerset Hospital *	100	50	9 (18.0)	28 (56.0)	13 (26.0)	50	10 (20.0)	27 (54.0)	13 (26.0)
Newark Beth Israel Medical Center	100	50	22 (44.0)	22 (44.0)	6 (12.0)	50	22 (44.0)	22 (44.0)	6 (12.0)
Osaka Metropolitan University	99	50	22 (44.0)	25 (50.0)	3 (6.0)	49	23 (46.9)	23 (46.9)	3 (6.1)
Puerta del Mar University Hospital	99	50	16 (32.0)	23 (46.0)	11 (22.0)	49	17 (34.7)	20 (40.8)	12 (24.5)
SES Hospital	98	50	27 (54.0)	15 (30.0)	8 (16.0)	48	24 (50.0)	16 (33.3)	8 (16.7)
Shiraz University of Medical Sciences	97	50	20 (40.0)	23 (46.0)	7 (14.0)	47	19 (40.4)	21 (44.7)	7 (14.9)
Sidra Medicine	98	50	30 (60.0)	19 (38.0)	1 (2.0)	48	26 (54.2)	20 (41.7)	2 (4.2)
Tel Aviv Medical Center	100	50	20 (40.0)	25 (50.0)	5 (10.0)	50	20 (40.0)	25 (50.0)	5 (10.0)
Tri-Service General Hospital	98	50	19 (38.0)	25 (50.0)	6 (12.0)	48	19 (39.6)	22 (45.8)	7 (14.6)
University Clinical Centre Kragujevac	100	50	20 (40.0)	26 (52.0)	4 (8.0)	50	21 (42.0)	24 (48.0)	5 (10.0)
University Hospital Aachen	98	50	17 (34.0)	24 (48.0)	9 (18.0)	48	17 (35.4)	21 (43.8)	10 (20.8)
University Hospital of Tuebingen	93	50	24 (48.0)	22 (44.0)	4 (8.0)	43	20 (46.5)	19 (44.2)	4 (9.3)
University of Alberta	98	50	7 (14.0)	31 (62.0)	12 (24.0)	48	7 (14.6)	28 (58.3)	13 (27.1)
University of Essen	94	50	26 (52.0)	21 (42.0)	3 (6.0)	44	20 (45.5)	20 (45.5)	4 (9.1)

Overview of the train and test set compositions across NICU sites. Low, normal, and high refer to the clinical labeling of the intubations. * New Somerset hospital only included for the exploratory training analysis looking at the impact of including scans with lower resolution within continual learning, as detailed in the Supplementary Methods.

Extended Data Table 2 | Demographic distribution between the aggregate train and test sets

Characteristic	Training Set n = 1,499	Test Set n = 1,461
Age, days		
No. with data (%)	1449 (96.7)	1408 (96.4)
Mean (SD)	16.7 (45.2)	14.5 (36.6)
Median (IQR)	2 (1-11)	2 (1-10)
Birth weight, g		
No. with data (%)	1339 (89.3)	1304 (89.3)
Mean (SD)	2179.5 (1242.6)	2209.9 (1225.4)
Median (IQR)	2100 (1132.5-2992.5)	2210 (1080-3030)
Gestational age, wk		
No. with data (%)	1297 (86.5)	1253 (85.8)
Mean (SD)	32.8 (5.4)	33 (5.3)
Median (IQR)	34 (28-37.9)	34 (28-38)
Sex, No. (%)		
Male	829 (55.3)	825 (56.5)
Female	666 (44.4)	625 (42.8)
Unknown	4 (0.3)	11 (0.8)
Race, No. (%)		
White	386 (25.8)	381 (26.1)
Black	39 (2.6)	39 (2.7)
Asian	477 (31.8)	466 (31.9)
American Indian	5 (0.3)	1 (0.1)
Pacific Islander	3 (0.2)	0 (0.0)
Other	47 (3.1)	48 (3.3)
Unknown	542 (36.2)	526 (36.0)

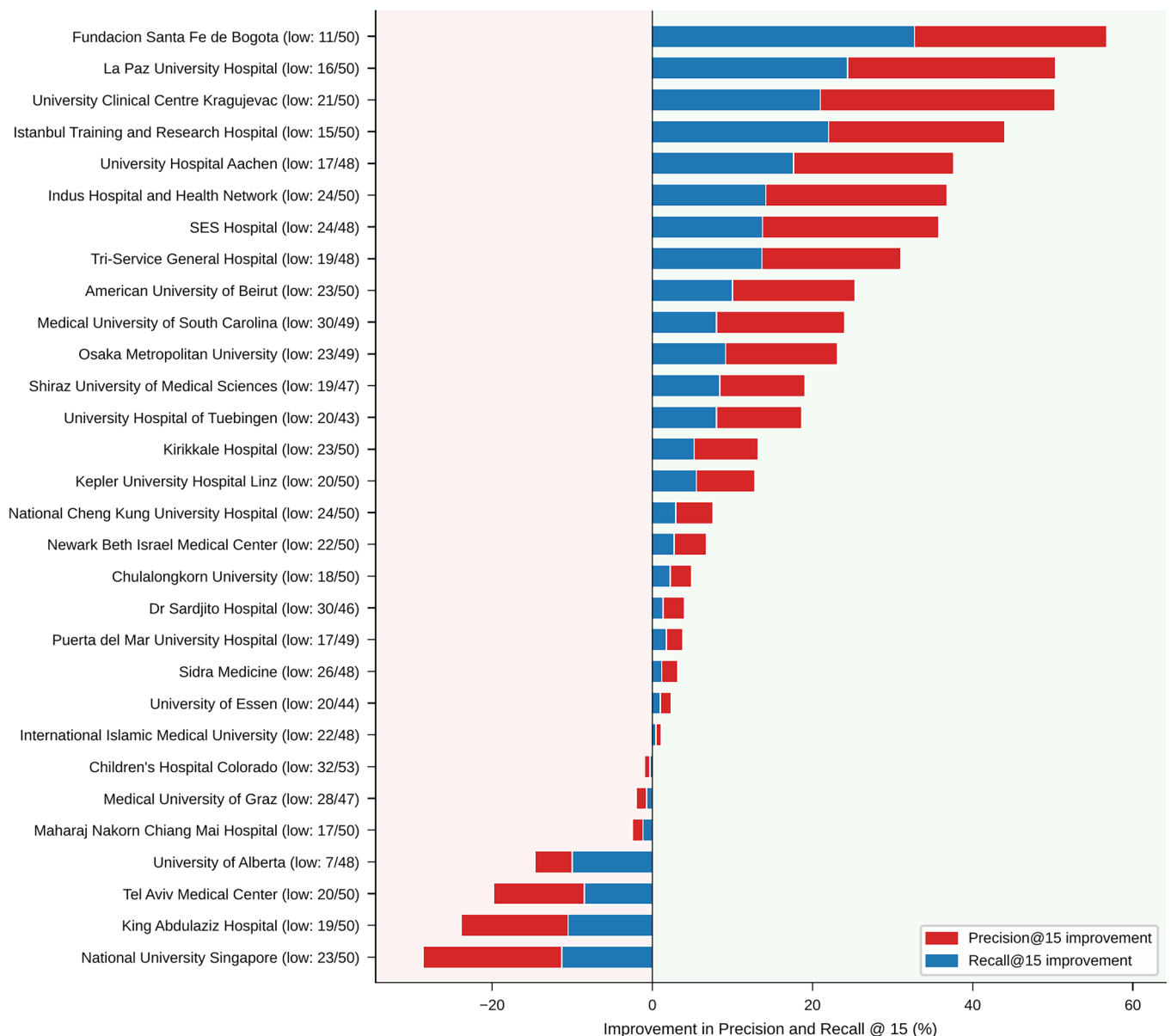
Demographic distribution between the aggregate train and test sets. Note that the total numbers here correspond to the number of unique patients that provided radiograph samples; 5 patients provided multiple samples, making the total training set and testing set 1500 and 1465 images respectively, as reflected in the main manuscript.



Extended Data Fig. 1 | Comparison between continual learning, fine-tuning, and direct prediction on ETT and carina localization across the 30 NICU sites.

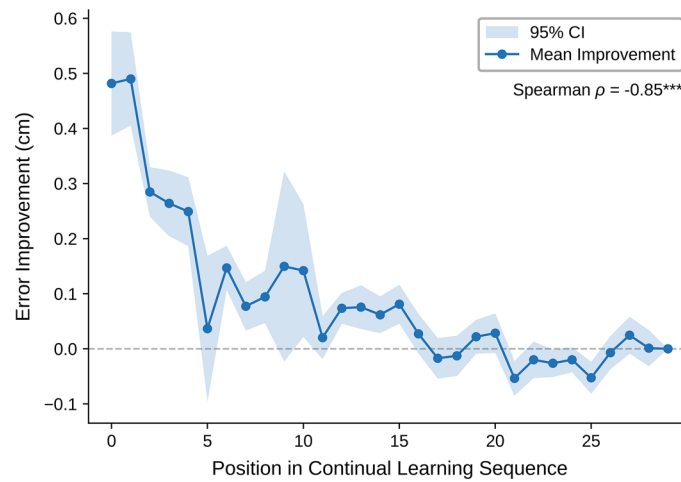
This figure includes the results of direct prediction, which are omitted in the main manuscript to better visualize the differences between continual learning and fine-tuning. Each row represents one NICU site ($n = 30$ sites). For continual learning and fine-tuning, the point shows the site's median combined ETT and carina error, averaged across 10 training simulations ($n = 10$ simulations per site);

random hospital orderings for continual learning, random training seeds for fine-tuning). For direct prediction (a single deterministic model), the point shows the median combined error across the site's test images. Dotted vertical lines indicate the mean of the per-site values across all 30 sites for each method. Error bars represent 95% confidence intervals across the 10 simulations for continual learning and fine-tuning, and bootstrap 95% confidence intervals ($B = 10,000$ resamples of test images) for direct prediction.



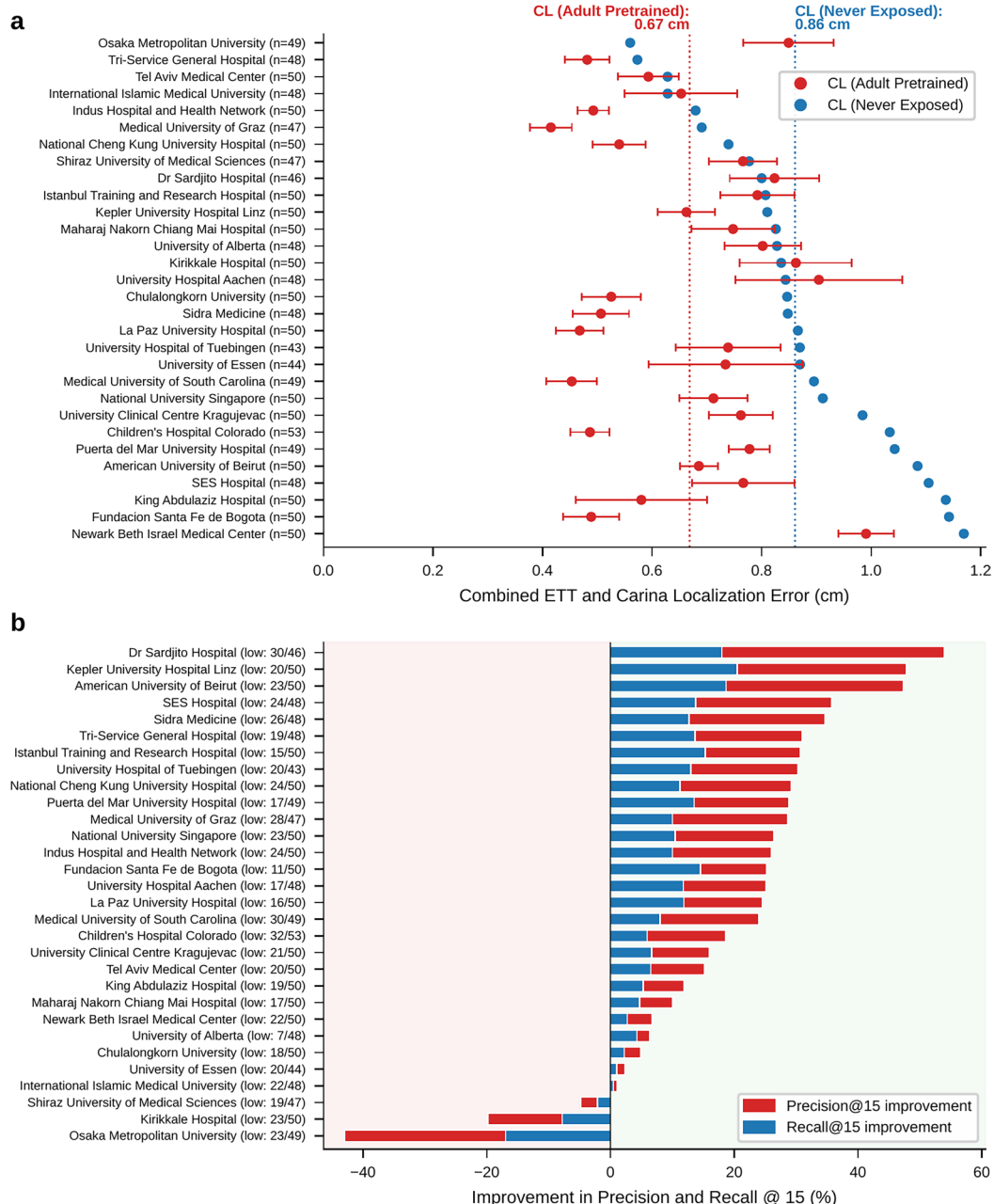
Extended Data Fig. 2 | Precision and recall@15 performance improvements for identifying abnormally low intubations when using continual learning compared to direct prediction. Each bar represents one of the 30 NICU sites and stacks the precision@15 and recall@15 improvements. For each metric, continual learning is averaged across its 10 simulations (random hospital

orderings) while direct prediction is a single deterministic model, and the bar shows the difference (continual learning - direct prediction). Sites are ordered by total improvement. Per-site 95% confidence intervals and significance are reported in Supplementary Tables 5-6.



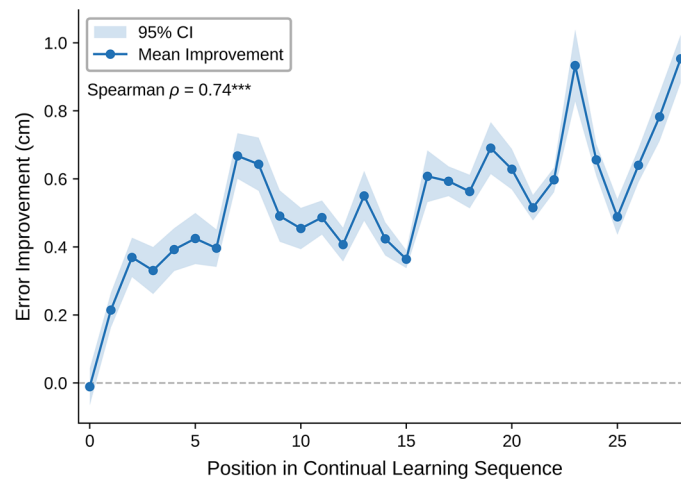
Extended Data Fig. 3 | Backward evaluation of catastrophic forgetting across the continual learning sequence. Each point represents the mean difference in combined ETT and carina localization error between the final model (after completing the full training sequence) and the intermediate model (evaluated immediately after training on the hospital at that position). Positive values

indicate that the final model outperforms the intermediate model. Results are averaged across 100 random hospital orderings ($n = 100$ orderings); the points indicate the mean difference and the shaded band is the 95% confidence interval across the 100 orderings.



Extended Data Fig. 4 | Continual learning with versus without adult ICU pretraining, evaluated on localization (a) and triage (b) across the 30 NICU sites. a. Per-site combined ETT and carina localization error for continual learning initialized from a model pretrained on adult ICU radiographs versus the same procedure initialized from a model never exposed to chest radiographs, across the 30 NICU sites (each row is one site). For the adult-pretrained condition, the point shows the site's median combined ETT and carina error averaged across 10 training simulations (random hospital orderings), with error bars showing 95% confidence intervals across the 10 simulations; the non-pretrained

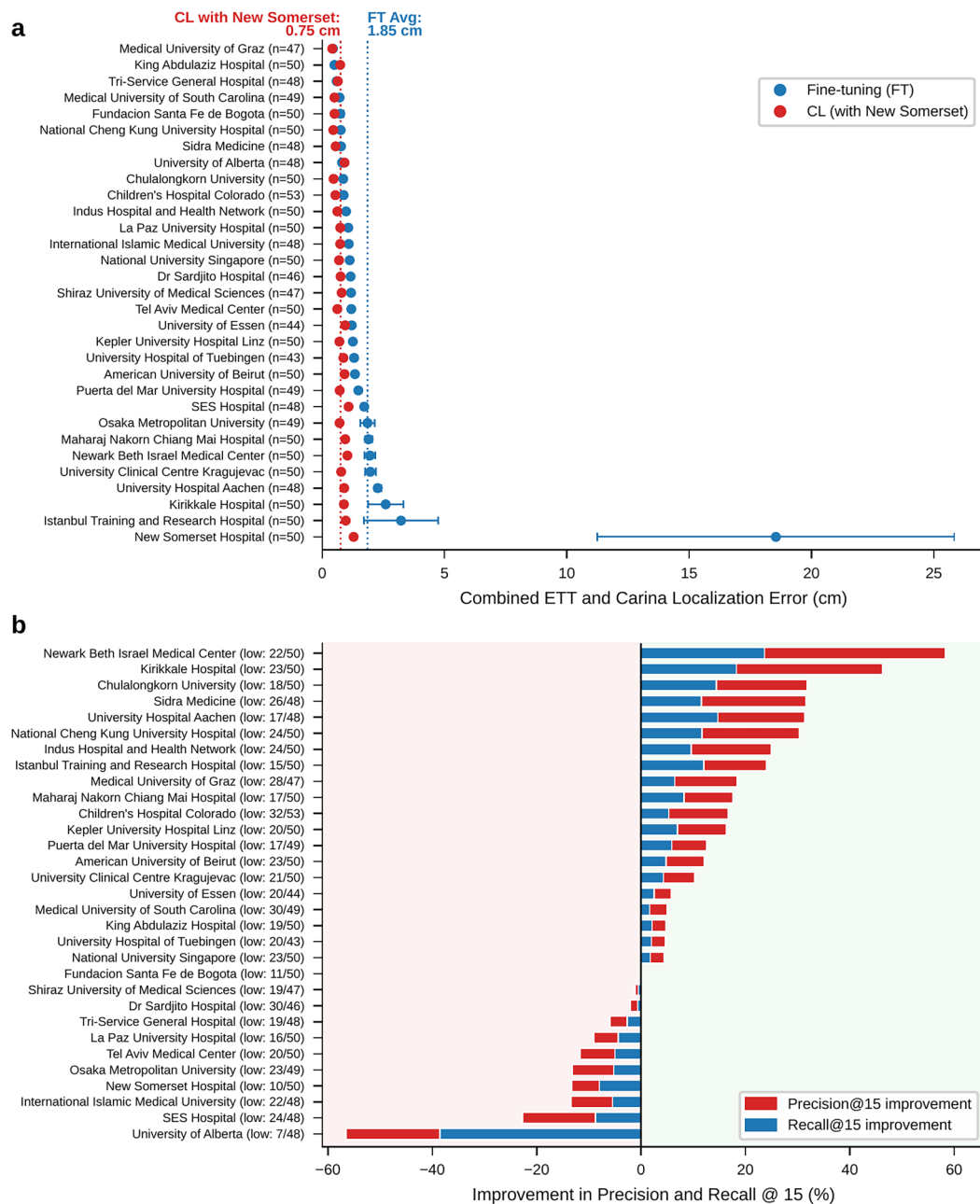
condition is a single training run per site. Dotted vertical lines indicate the mean of the per-site values across all 30 sites for each condition. **b.** Improvement conferred by adult pretraining on identifying the most concerning scans for radiologist confirmation. Each bar represents one of the 30 NICU sites and stacks the precision@15 and recall@15 improvements for detecting abnormally low intubations. For each metric, the adult-pretrained condition is averaged across its 10 simulations (random hospital orderings) while the non-pretrained condition is a single run, and the bar shows the difference (adult-pretrained - non-pretrained). Sites are ordered by total improvement.



Extended Data Fig. 5 | Improvement of combined carina and ETT tip localization based on hospital position in the continual learning sequence.

Each point represents the average improvement in performance for a hospital at a given position in the sequence versus training the model only on data from the site in that position. The improvements for each position are pooled across all

holdout-site $\times$ ordering combinations ($n = 300$ per position; 30 sites $\times$ 10 random orderings); the point is the mean and the band is the 95% confidence interval across these orderings. The increase in improvement with sequence position is quantified by a two-sided Spearman rank correlation across the 29 sequence positions ($\rho = 0.74$, $p < 0.001$).



Extended Data Fig. 6 | Continual learning including the omitted (New Somerset) hospital versus fine-tuning across the 31 NICU sites, for localization (a) and triage (b). **a.** Per-site combined ETT and carina localization error for continual learning performed with the omitted hospital's low-resolution scans included in the training sequence versus fine-tuning, across the 31 NICU sites (each row is one site). For fine-tuning, the point shows the site's median combined ETT and carina error, averaged across 10 training simulations (random training seeds), with error bars showing 95% confidence intervals across the 10 simulations; the continual-learning results including the omitted hospital are from a single hospital ordering per site. Dotted vertical lines indicate the mean

of the per-site values across all 31 sites for each condition. **b.** Improvement conferred by continual learning using the omitted hospital versus fine-tuning on identifying the most concerning scans for radiologist confirmation. Each bar represents one of the 31 NICU sites and stacks the precision@15 and recall@15 improvements for detecting abnormally low intubations. For each metric, fine-tuning is averaged across its 10 simulations (random training seeds) while continual learning with the omitted hospital is a single ordering, and the bar shows the difference (continual learning - fine-tuning). Sites are ordered by total improvement.

Reporting Summary

Nature Portfolio wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Portfolio policies, see our [Editorial Policies](#) and the [Editorial Policy Checklist](#).

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a Confirmed

- ☐ ☒ The exact sample size (n) for each experimental group/condition, given as a discrete number and unit of measurement
- ☐ ☒ A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
- ☐ ☒ The statistical test(s) used AND whether they are one- or two-sided
Only common tests should be described solely by name; describe more complex techniques in the Methods section.
- ☐ ☒ A description of all covariates tested
- ☐ ☒ A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
- ☐ ☒ A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
- ☐ ☒ For null hypothesis testing, the test statistic (e.g. F , t , r) with confidence intervals, effect sizes, degrees of freedom and P value noted
Give P values as exact values whenever suitable.
- ☒ ☐ For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
- ☒ ☐ For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
- ☒ ☐ Estimates of effect sizes (e.g. Cohen's d , Pearson's r), indicating how they were calculated

Our web collection on [statistics for biologists](#) contains articles on many of the points above.

Software and code

Policy information about [availability of computer code](#)

Data collection The Redbrick commercial data annotation platform was used.

Data analysis Analyses were performed in Python 3.9.19. Model development and inference used PyTorch 2.6.0 and torchvision 0.21.0 (CarinaNet, a RetinaNet-based architecture adapted from Oliver et al. 2023). Image preprocessing used scikit-image 0.24.0 and OpenCV 4.11.0.86. Statistical analyses used SciPy 1.13.1, statsmodels 0.14.2, and scikit-learn 1.6.1. Data handling used NumPy 2.0.2 and pandas 2.2.2; figures were generated with matplotlib 3.9.4 and seaborn 0.13.2. The complete computational environment is specified in the environment file in the source code repository.
Source code: <https://github.com/rajpurkarlab/MAIDA-NICU-continual-learning>

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Portfolio [guidelines for submitting code & software](#) for further information.

Data

Policy information about [availability of data](#)

All manuscripts must include a [data availability statement](#). This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A description of any restrictions on data availability
- For clinical datasets or third party data, please ensure that the statement adheres to our [policy](#)

The chest radiograph images that support the findings of this study are currently not publicly available because the data use agreements governing this multi-institutional international collaboration prohibit the transfer or redistribution of patient radiographs and patient privacy protections preclude their release. Aggregate summary statistics and demographic characteristics necessary to interpret the findings are provided in the Extended Data Tables and Supplementary Information. The trained model weights, which enable application of the continual learning framework to new datasets, are publicly available at <https://github.com/rajpurkarlab/MAIDA-NICU-continual-learning>.

Research involving human participants, their data, or biological material

Policy information about studies with [human participants or human data](#). See also policy information about [sex, gender \(identity/presentation\), and sexual orientation](#) and [race, ethnicity and racism](#).

Reporting on sex and gender

Sex was recorded based on clinical documentation at each participating institution. The cohort included 55.9% male (n=1,654), 43.6% female (n=1,291), and 0.5% unknown or unreported (n=15). The distribution of the dataset based on sex was reported in more detail in the supplement. Sex-stratified analyses were performed to evaluate potential disparities in model performance, with results reported in Supplemental Tables 3-4. Sex was not used as a variable in model training or prediction.

Reporting on race, ethnicity, or other socially relevant groupings

Race and ethnicity were recorded based on clinical documentation at each participating institution using locally defined categories. The cohort included Asian (31.9%, n=943), White (25.9%, n=767), Other (3.2%, n=95), Black or African American (2.6%, n=78), American Indian/Alaska Native (0.2%, n=6), Pacific Islander (0.1%, n=3), and unreported (36.1%, n=1,068). Race/ethnicity-stratified analyses were performed to evaluate potential disparities in model performance shown in Supplemental Tables 3-4.

Population characteristics

This international multi-center study included 2,965 NICU chest radiographs from 2,960 unique patients across 30 participating centers in 20 countries and territories. The median age at radiography was 2 days (IQR, 1-11 days). The mean birth weight was 2,195 grams (SD, 1,234 grams) and mean gestational age was 32.9 weeks (SD, 5.3 weeks), reflecting a predominantly preterm population. Male infants comprised 55.9% of the cohort. Full demographic characteristics stratified by site and train/test split are provided in Extended Data Tables 1-2 and Supplemental Table 1.

Recruitment

This was a retrospective study of existing chest radiographs. Eligible patients were NICU patients aged 0-6 months who underwent bedside chest radiography with an endotracheal tube in place. Each participating institution collected approximately 100 radiographs, with samples selected to provide balanced representation of ETT positioning (approximately 50% correctly positioned, 50% incorrectly positioned) as determined by radiologist interpretation. Samples were primarily collected from 2021, with earlier years included when insufficient cases were available. The class-balanced design was implemented to ensure adequate representation of abnormal tube placements for model training and evaluation. However, these datasets as a result may not be representative of the true clinical prevalence of abnormal endotracheal placements.

Ethics oversight

This study was approved by the Harvard Longwood Campus Institutional Review Board (HLC IRB) at Harvard Medical School (IRB22-0400). Each participating institution obtained local institutional review board approval for data collection and sharing. Patients or the public were not involved in the design, conduct, reporting, or dissemination of this research.

Note that full information on the approval of the study protocol must also be provided in the manuscript.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see [nature.com/documents/nr-reporting-summary-flat.pdf](https://www.nature.com/documents/nr-reporting-summary-flat.pdf)

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size

Each participating hospital contributed approximately 100 chest radiographs, with 50 allocated to training and the remainder to testing. This sample size was determined by consensus among collaborating institutions as a modest, feasible contribution that could be reliably collected across diverse international settings with varying resource constraints. No formal power calculation was performed; however, the multi-site design with 30 hospitals provided sufficient data to evaluate model generalization across heterogeneous clinical environments.

Data exclusions

New Somerset Hospital data were excluded from the primary analysis due to substantially lower radiograph resolution compared to other

Data exclusions	sites; these data were included in exploratory analyses evaluating the impact of image quality on continual learning performance (Extended Data Fig. 6).
Replication	To account for potential ordering effects in the continual learning sequence, results were averaged across 10 random hospital orderings for each continual learning model. Performance improvement by sequence position was averaged across 300 held-out-site and ordering combinations (30 sites × 10 random orderings). The consistency of improvements across the majority of test sites taken across all of these simulations supports the reproducibility of the findings.
Randomization	Train and test splits were stratified by ETT placement classification (low, normal, or high) to maintain consistent label distributions across both sets (within those stratifications, images were randomly selected). Images were randomly sampled for annotation to minimize systematic bias. The order of hospitals in the continual learning sequence was randomized across multiple trials to evaluate ordering effects.
Blinding	Images were annotated by medical students, reviewed by a radiology resident and a physician, and a board-certified radiologist verified a random 20% of samples; annotators were blinded to model predictions. Model evaluation was performed using held-out test sets that were not used during any stage of model development or hyperparameter tuning.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☒	☐ Antibodies
☒	☐ Eukaryotic cell lines
☒	☐ Palaeontology and archaeology
☒	☐ Animals and other organisms
☐	☒ Clinical data
☒	☐ Dual use research of concern
☒	☐ Plants

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☒	☐ MRI-based neuroimaging

Clinical data

Policy information about [clinical studies](#)

All manuscripts should comply with the ICMJE [guidelines for publication of clinical research](#) and a completed [CONSORT checklist](#) must be included with all submissions.

Clinical trial registration	This was a retrospective observational study (not a randomized trial) and was not registered as a clinical trial.
Study protocol	The study protocol was approved by the Harvard Longwood Campus Institutional Review Board (HLC IRB) at Harvard Medical School (IRB22-0400). Each participating institution followed a standardized data collection protocol specifying inclusion criteria, annotation procedures, and de-identification requirements.
Data collection	Chest radiographs were retrospectively collected from 30 NICUs across 20 countries and territories. Samples were primarily collected from 2021, with earlier years included when insufficient cases were available. For each case, radiograph images, technical imaging parameters, patient demographics (sex, gestational age, birth weight), and clinical indication were collected. Expert annotations of ETT tip and carina positions were performed using standardized bounding box protocols on the Redbrick annotation platform.
Outcomes	The primary outcome was anatomical landmark localization error, measured as the Euclidean distance (in centimeters) between model-predicted and ground-truth positions for the ETT tip and carina. Secondary outcomes included precision@k and recall@k for identifying deep intubations among the k most concerning predictions, and false positive rates. Clinical relevance was assessed by evaluating whether localization improvements corresponded with improved triaging of abnormally deep ETT placements.

Plants

Seed stocks	n/a
Novel plant genotypes	n/a
Authentication	n/a