THE UNIVERSITY of ADELAIDE

Closing the implementation gap in pre-deployment medical AI study design

Dr Luke Oakden-Rayner, MBBS FRANZCR

Submitted November 2021

A thesis by publication submitted in fulfilment of the requirements for the degree of Doctor of Philosophy in the School of Public Health, University of Adelaide.

Table of Contents

Publications comprising this thesis	4
Other publications arising from PhD project but not forming part of this thesis	5-6
Abstract	7
Declaration	8
General acknowledgements	9
Introduction	10
Section 1: Issues with medical data	11
1.1 Data labels	12-20
Paper: *Exploring large-scale public medical imaging datasets*	
1.2 Data subsets	21-31
Paper: *Hidden stratification causes clinically significant failures in machine learning for medical imaging*	
Section 2: Issues with testing	32
2.1 Measuring human performance	33-48
Paper: *Docs are ROCs: A simple off-the-shelf approach for estimating average human performance in diagnostic studies*	
2.2 Error analysis	49-84
Paper: *The medical algorithmic audit*	
Section 3: Explainability	85-92
Paper: *The false hope of explainable AI*	

Section 4: Application of the lessons 93-126

Paper: *Validation and algorithmic audit of a deep learning system*

for the detection of hip fractures in emergency department patients

Conclusions 127

Bibliography 128-132

Publications comprising this thesis

1. **Luke Oakden-Rayner**. "Exploring large-scale public medical image datasets." *Academic radiology* 27.1 (2020): 106-112.
2. **Luke Oakden-Rayner**, Jared Dunnmon, Gustavo Carneiro, Christopher Ré. "Hidden stratification causes clinically meaningful failures in machine learning for medical imaging." *Proceedings of the ACM conference on health, inference, and learning*. 2020.
3. **Luke Oakden-Rayner**, and Lyle J. Palmer. "Docs are ROCs: a simple off-the-shelf approach for estimating average human performance in diagnostic studies." *arXiv preprint arXiv:2009.11060* (2020).
4. Xiaoxuan Liu, Ben Glocker, Melissa McCradden, Marzyeh Ghassemi, Alastair K. Denniston, **Luke Oakden-Rayner**. "The Medical Algorithmic Audit." *Lancet Digital Health (in press)*. 2021.
5. Ghassemi, Marzyeh, **Luke Oakden-Rayner**, and Andrew L. Beam. "The false hope of current approaches to explainable artificial intelligence in health care." *The Lancet Digital Health* 3.11 (2021): e745-e750.
6. **Oakden-Rayner, Luke**, William Gale, Thomas A. Bonham, Matthew P. Lungren, Gustavo Carneiro, Andrew P. Bradley, Lyle J. Palmer. "Validation and algorithmic audit of a deep learning system for the detection of proximal femoral fractures in emergency department patients: a diagnostic accuracy study." *Lancet Digital Health (in press)*. 2021.

Other publications arising from PhD project but not forming part of this thesis

1. Gustavo Carneiro, **Luke Oakden-Rayner**, et al. "Automated 5-year mortality prediction using deep learning and radiomics features from chest computed tomography." *2017 IEEE 14th International Symposium on Biomedical Imaging (ISBI 2017)*. IEEE, 2017.
2. **Luke Oakden-Rayner**, et al. "Precision radiology: predicting longevity using feature engineering and deep learning methods in a radiomics framework." *Scientific reports* 7.1 (2017): 1-13.
3. William Gale, **Luke Oakden-Rayner**, et al. "Detecting hip fractures with radiologist-level performance using deep neural networks." *arXiv preprint arXiv:1711.06504* (2017).
4. **Luke Oakden-Rayner**, Andrew L. Beam, and Lyle J. Palmer. "Medical journals should embrace preprints to address the reproducibility crisis." (2018): 1363-1365.
5. Samuel G. Finlayson, Hyunkwang Lee, Isaac S. Kohane, **Luke Oakden-Rayner**. "Towards generative adversarial networks as a new paradigm for radiology education." *arXiv preprint arXiv:1812.01547* (2018).
6. **Luke Oakden-Rayner**, and Lyle John Palmer. "Artificial intelligence in medicine: validation and study design." *Artificial Intelligence in Medical Imaging*. Springer, Cham, 2019. 83-104.
7. Stephen Bacchi, **Luke Oakden-Rayner**, et al. "Deep learning natural language processing successfully predicts the cerebrovascular cause of transient ischemic attack-like presentations." *Stroke* 50.3 (2019): 758-760.
8. William Gale, **Luke Oakden-Rayner**, et al. "Producing Radiologist-Quality Reports for Interpretable Deep Learning." *2019 IEEE 16th international symposium on biomedical imaging (ISBI 2019)*. IEEE, 2019.
9. Marcus A. Badgeley, John R. Zech, **Luke Oakden-Rayner**, et al. "Deep learning predicts hip fracture using confounding patient and healthcare variables." *NPJ digital medicine* 2.1 (2019): 1-10.
10. **Luke Oakden-Rayner**. "The Rebirth of CAD: How Is Modern AI Different from the CAD We Know?." *Radiology. Artificial intelligence* 1.3 (2019).
11. Stephen Bacchi, …, **Luke Oakden-Rayner**, Sandy Patel. "Deep learning in the detection of high-grade glioma recurrence using multiple MRI sequences: a pilot study." *Journal of Clinical Neuroscience* 70 (2019): 11-13.
12. Stephen Bacchi, Toby Zerner, **Luke Oakden-Rayner**, et al. "Deep learning in the prediction of ischaemic stroke thrombolysis functional outcomes: a pilot study." *Academic radiology* 27.2 (2020): e19-e23.
13. Stephen Bacchi, **Luke Oakden-Rayner**, et al. "Stroke prognostication for discharge planning with machine learning: A derivation study." *Journal of Clinical Neuroscience* 79 (2020): 100-103.
14. Samantha Cruz Rivera, ..., **Luke Oakden-Rayner**, et al. "Guidelines for clinical trial protocols for interventions involving artificial intelligence: the SPIRIT-AI extension." *Nature Medicine* 26.9 (2020): 1351-1363.
15. Xiaoxuan Liu, ..., **Luke Oakden-Rayner**, et al. "Reporting guidelines for clinical trial reports for interventions involving artificial intelligence: the CONSORT-AI extension." *bmj* 370 (2020).
16. Stephen Bacchi, Yiran Tan, **Luke Oakden-Rayner**, et al. "Machine Learning in the Prediction of Medical Inpatient Length of Stay." *Internal Medicine Journal* (2020).

17. Hugh Harvey and **Luke Oakden-Rayner**. "Guidance for Interventional Trials Involving Artificial Intelligence." *Radiology: Artificial Intelligence* 2.6 (2020): e200228.
18. James John Joseph Condon, **Luke Oakden-Rayner**, et al. "Replication of an open-access deep learning system for screening mammography: Reduced performance mitigated by retraining on local data." *medRxiv* (2021).
19. Jane Scheetz, …, **Luke Oakden-Rayner**, et al. "A survey of clinicians on the use of artificial intelligence in ophthalmology, dermatology, radiology and radiation oncology." *Scientific reports* 11.1 (2021): 1-10.
20. Felix Paterson, ..., **Luke Oakden-Rayner**. "Assessing the accuracy of 68Ga-PSMA PET/CT compared with MRI in the initial diagnosis of prostate malignancy: A cohort analysis of 114 consecutive patients." *Journal of Medical Imaging and Radiation Oncology* (2021).
21. Banerjee, Imon, …, **Luke Oakden-Rayner**, et al. "Reading Race: AI Recognises Patient's Racial Identity In Medical Images." *arXiv preprint arXiv:2107.10356* (2021).
22. Catherine M. Jones, Quinlan D. Buchlak, **Luke Oakden-Rayner**, et al. "Chest radiographs and machine learning–Past, present and future." *Journal of Medical Imaging and Radiation Oncology* 65.5 (2021): 538-544.
23. Jarrel CY Seah, …, **Luke Oakden-Rayner**, et al. "Effect of a comprehensive deep-learning model on the accuracy of chest x-ray interpretation by radiologists: a retrospective, multireader multicase study." *The Lancet Digital Health* 3.8 (2021): e496-e506.
24. Charlotte Blacketer, …, **Luke Oakden-Rayner**, et al. "Medical student knowledge and critical appraisal of machine learning: a multicentre international cross-sectional study." *Internal medicine journal* 51.9 (2021): 1539-1542.

Abstract

The rapid development of clinical artificial intelligence (AI) technologies has outpaced the development of robust regulatory and clinical safety mechanisms. AI systems are cleared for use and deployed in practice relying on pre-clinical performance studies, without evidence of the impact this will have on patient and provider outcomes. This has led to concerns of an “implementation gap”, where systems that appear to perform well on pre-clinical testing fail to produce the expected outcomes in practice.

While there is an urgent need for direct clinical testing of AI systems and evaluation of the impact of these systems on patient and provider outcomes, it is implausible to expect the clinical evaluation will be performed at the scale necessary to mitigate potential AI harms of the many AI systems already in use and currently under development.

In this body of work I look at factors which may contribute to the implementation gap, in particular the effects of low-quality training and testing data, flawed and incomplete study design methodologies, and an over-reliance on explainability methods to address safety. I suggest a series of improvements to how we design, evaluate, and utilise AI systems in clinical practice, with the goal of better estimating the potential harms of AI during the pre-clinical testing phase, and by doing so closing the implementation gap.

Declaration

I certify that this work contains no material which has been accepted for the award of any other degree or diploma in my name in any university or other tertiary institution and, to the best of my knowledge and belief, contains no material previously published or written by another person, except where due reference has been made in the text. In addition, I certify that no part of this work will, in the future, be used in a submission in my name for any other degree or diploma in any university or other tertiary institution without the prior approval of the University of Adelaide and where applicable, any partner institution responsible for the joint award of this degree.

The author acknowledges that copyright of published works contained within this thesis resides with the copyright holder(s) of those works.

I give permission for the digital version of my thesis to be made available on the web, via the University's digital research repository, the Library Search and through web search engines, unless permission has been granted by the University to restrict access for a period of time.

I acknowledge the support I have received for my research through the provision of an Australian Government Research Training Program Scholarship.

Luke Oakden-Rayner

Acknowledgements

I would like to acknowledge the many people who have helped me during my PhD candidature, without whom this thesis would not have been possible.

First, to my partner Anne, who has been an unlimited source of support and kindness during my candidature, and to our children Alma and Drew. You have always been my moral compass, my heart, and my inspiration. You make me want to be a better researcher and a better person. To my wider family, and in particular my brother and mum, thank you for all of your support.

To my supervisor and mentor Lyle Palmer; I would not have embarked on this journey without your encouragement, and I could not have succeeded without your patience, generosity, guidance, or friendship.

To my other supervisors Gustavo Carneiro and Taryn Bessen, I am grateful for all of your advice and valued contributions over these years.

To my colleagues, collaborators, and friends, I thank you all for your support and assistance: Andrew Bradley, Jared Dunnmon, Xiaoxuan Liu, Andrew Beam, Marzyeh Ghassemi, Alastair Denniston, Gabriel Maicas, Pearse Keane, Sam Finlayson, Judy Gichoya, Matt Lungren, Curt Langlotz, Chris Ré, William Gale, Stephen Bacchi, Christopher Kelly, Toby Zerner, Minh-Son To, Zak Kohane, Catherine Jones, Quin Buchlak, Jarrel Seah, Lily Peng, Hugh Harvey, Chuck Kahn, Erik Ranschaert, Melissa McCradden, Luke Smith, Minyan Zeng, James Condon, Ryan Pham, Brandon Price, John Zech, Marcus Badgeley, and Johan Verjans. I'm sorry for those I have missed.

I would like to acknowledge those who have inspired me to think about my own work in a different way. There are too many to recognise individually, but in particular I would like to note the work of Timnit Gebru and Margaret Mitchell, Katherine Heller and Alexander D'Amour, Deb Raji, Andrew Selbst, and Robert Geirhos. Each of you has profoundly changed my understanding of AI.

Finally, I also want to acknowledge the support and kindness I have received from the broader community more recently. Coming out as transgender during a PhD and as an early career researcher has been daunting. Despite my fears, the love and acceptance I have received has been wonderful and beyond my most optimistic expectations. As well as those already acknowledged and the many others who have expressed their kind wishes and congratulations, I would like to specifically mention the transgender and queer researchers whose simple act of being visible has helped me enormously. Again, there are too many to note, but a few in particular have been important to me. To Danielle Navarro, you taught me that transitioning as an academic in South Australia was even possible. To Alex Hanna, you showed me that we don't need to hide parts of ourselves from the world but instead can carry ourselves with pride. To my friend and colleague Alix Bird, who carved out the path that I have followed, you inspired me to be *me*; to live my life unbound by social norms and to question the ideas and beliefs I hold through inertia instead of intention. And to all of the wonderful people who are involved in Queer In AI, who showed me that I am not alone.

With much love and gratitude,

Lauren :)

Introduction: The AI Implementation Gap

The explosion of artificial intelligence (AI) research [1,2] and commercial ventures [3] since the technological breakthroughs of the early 2010s has left regulators and governance bodies struggling to catch up [3,4]. Indeed, dozens of medical AI systems have been approved by the FDA in the last decade [5], despite an absence of regulations for software as a medical device (SaMD) until December 2017 [6], and an ongoing lack of AI-specific regulatory guidance (a document outlining the future plans of the regulator regarding medical AI was released in early 2021 [7]).

In this environment, many have raised concerns about an "implementation gap" in medical AI, where promising preclinical AI models may fail to translate into actionable, safe, and efficacious clinical tools [8]. There is historical precedent that suggests the gap may be caused, at least in part, by a failure of pre-clinical testing, where the exciting preclinical performance results failed to predict the poor clinical performance of the deployed models. This phenomenon was observed with an older form of AI technology broadly termed computer-aided detection (CAD) algorithms. These systems were widely deployed in screening mammography programs in the USA during the 1990s and 2000s [9], but the positive preclinical evidence did not translate into clinical benefit, and may have in fact caused clinical harm [9,10].

More recently, a growing body of clinical and research experience with medical AI has validated these concerns, with many examples of AI models failing in unexpected and often potentially harmful ways, including models which rely on spurious image features [11–13], models which exacerbate existing healthcare disparities [14,15], and models which tend to underperform when applied to new populations (external validation) [1].

The obvious solution is to test these systems *clinically*, to demonstrate that not only is performance similar to human experts in controlled laboratory conditions, but that patient and healthcare outcomes are equal or improved in the real world [16]. The gold standard approach to clinical testing is the interventional randomised control trial (RCT), where the AI model is deployed clinically and closely monitored, much like a pharmaceutical trial. Recently, guidelines for the design and reporting of clinical trials for medical AI have been developed [17,18].

There can be little to no implementation gap with an RCT as it directly measures the impact of implementation. However, RCTs are expensive, slow, and can be commercially risky [16]. The FDA has taken the pragmatic stance that low and medium risk AI systems [19] can be safely sold without interventional RCTs, albeit with the requirement to perform post-marketing monitoring [6]. The safety of this approach is yet to be tested, as thus far no "high-risk" systems have undergone regulatory review and few reported studies have included prospective randomisation in their design [2,20].

The conflict between safety and pragmatism in medical AI safety is the foundation of this thesis and body of work. If we accept that many AI systems will not undergo real-world clinical evaluation prior to marketing, is there a way to **close the implementation gap** by augmenting preclinical testing? As a general approach, I consider what factors may be responsible for the implementation gap, and rather than seeking technical solutions, I ask if there are potential human-lead strategies to produce more informative preclinical results and to identify AI model vulnerabilities?

Section 1: Issues with medical data

Machine learning applied to medical problems, similar to much of clinical epidemiology and medical research in general, is a data driven science. Both scale and quality of data are therefore critical aspects in constructing robust and useful clinical AI systems.

Training data is often described as one of the most critical parts of any AI development process, and is commonly implicated as a cause of the unexpectedly poor real world performance, unfairness, and algorithmic injustice [21].

Most often, this is discussed in terms of the size of datasets, with an expectation of "more is better". Certainly, it has been shown that AI performance improves as data increases, with no known ceiling [22], although extremely large datasets can be unwieldy and difficult to operate on. Certainly, in other fields AI performance continues to improve even into hundreds of thousands or millions of examples.

In medicine, data availability is heavily limited by disease rarity - even common diseases often occur only thousands of times per year in a given hospital or healthcare network, and uncommon diseases can be much rarer. This leads to highly imbalanced datasets, where there may be hundreds of thousands of cases, but only hundreds or even tens of examples of a particular pathology and tending to reflect only the most common variants of the disease. Take for example screening mammography, with a detected breast cancer prevalence of 0.3-0.7%, or less than one cancer per 100 disease-free patients [23]. The implication is that for many medical problems, it can be incredibly difficult or even impossible to obtain "enough" data.

Not just the *quantity* of data but also the *quality* of data is extremely important in medical AI, as Gebru et al say: "the characteristics of these datasets will fundamentally influence a model's behavior" [21]. In this context, an extremely important but less discussed consideration is the quality of data labels. While it has been argued that deep learning models tend to be robust to label noise [24], this tends to relate to random noise, and both class-dependent label noise (where the label errors depend on the label class) and feature-dependent label noise (where the magnitude and type of the label errors are dependent on the content of the input data) cause more severe degradation [25]. In the medical context, sources of these types of structured noise are plentiful.

Finally, the effect of the visual conspicuity of features has been poorly investigated in the computer vision and medical imaging AI literature. A large proportion of the broader computer vision AI literature is benchmarked on visual tasks which contain highly conspicuous subjects; the features of interest are easily distinguished from surrounding structures and those features fill a large portion of the input images. In medical imaging, many features of interest are small and subtle and it is possible that this distinction may limit our ability to extrapolate findings from the broader computer vision literature directly to medical contexts.

1.1 The nature, quality, and content of input data labels

The choice of label categories and the labelling methods employed can directly result in an 'implementation gap'. Commonly, AI models are tested on a dataset sampled from the same source as the data used for training. Even when tested on external (independent) data, the labelling schema and labelling methods (these might be called the "label generating process") are often the same as the local data. If this label generating process is poorly matched to the intended purpose and use of an AI model, then not only will the model perform poorly when deployed, but it will perform unreasonably well when tested (as the errors in the labels are equally present in the test data).

An example of this problem is the use of natural language processing to label data in several large public chest x-ray datasets, notably CXR-14 [26], Chexpert [27], and MIMIC-CXR [28]. While the latter includes the free text radiology reports in the dataset, all of these datasets make the same underlying assumption; that the visual information in the image is adequately reflected in the reports to produce meaningful labels (although the Chexpert dataset does also include a small test set with labels produced by visual evaluation of the images by radiologists).

In "**Exploring Large-Scale Public Medical Image Datasets"** [29], I consider the accuracy of labels for the CXR-14 dataset [26] and the MURA (musculoskeletal radiographs) dataset [30], showing that the method of labelling has a direct impact on the reliability of the labels when compared to a gold standard for the task (expert human visual review of the images).

Exploring Large-Scale Public Medical Image Datasets

Publication status: Published in the journal *Academic Radiology, 2020*.

Contribution: This was a single author publication, 100%

Detailed description contribution: I planned and designed the study, performed the experiments, and drafted and edited the manuscript.

Certification from co-authors: Nil

Declaration: This publication was part of the work undertaken during and for my HDR candidature.

Original Investigation

Exploring Large-scale Public Medical Image Datasets

Luke Oakden-Rayner, MBBS, FRANZCR

Abbreviations

AI
artificial intelligence

CXR14
The ChestXray14 dataset

MURA
The Musculoskeletal Radiology dataset

RSNA
Radiological Society of North America

Rationale and Objectives: Medical artificial intelligence systems are dependent on well characterized large-scale datasets. Recently released public datasets have been of great interest to the field, but pose specific challenges due to the disconnect they cause between data generation and data usage, potentially limiting the utility of these datasets.

Materials and Methods: We visually explore two large public datasets, to determine how accurate the provided labels are and whether other subtle problems exist. The ChestXray14 dataset contains 112,120 frontal chest films, and the Musculoskeletal Radiology (MURA) dataset contains 40,561 upper limb radiographs. A subset of around 700 images from both datasets was reviewed by a board-certified radiologist, and the quality of the original labels was determined.

Results: The ChestXray14 labels did not accurately reflect the visual content of the images, with positive predictive values mostly between 10% and 30% lower than the values presented in the original documentation. There were other significant problems, with examples of hidden stratification and label disambiguation failure. The MURA labels were more accurate, but the original normal/abnormal labels were inaccurate for the subset of cases with degenerative joint disease, with a sensitivity of 60% and a specificity of 82%.

Conclusion: Visual inspection of images is a necessary component of understanding large image datasets. We recommend that teams producing public datasets should perform this important quality control procedure and include a thorough description of their findings, along with an explanation of the data generating procedures and labeling rules, in the documentation for their datasets.

Key Words: Artificial intelligence; dataset; exploratory analysis; deep learning; quality control.

INTRODUCTION

The successful training of modern artificial intelligence (AI) relies on large, well-characterized datasets (1). The availability of these datasets can be considered a major barrier to the production of high quality image analysis AI systems in radiology, not only because the cost to produce these datasets is high, but also because access to existing datasets is restricted. Privacy concerns around the sharing of medical data and the competitive advantage that medical AI companies obtain from their own proprietary datasets is likely to have limited the sharing of these resources.

To overcome this challenge, several large public datasets have been made available in recent years. The ChestXray14 (CXR14) dataset produced by a team of researchers at the National Institutes of Health Clinical Center contains over 112,000 chest radiographs (2). The Musculoskeletal Radiology (MURA) dataset and competition from the Stanford Machine Learning Group contains over 40,000 upper limb radiographs (3). The Radiological Society of North America (RSNA) Paediatric Bone Age challenge dataset contains 14,236 upper limb radiographs (4). Several other notable recent releases of data include the RSNA pneumonia challenge, which builds on the CXR14 dataset with radiologist-produced labels, the QC500 dataset from Qure.AI (a commercial group) containing 500 CT head images in patients with and without intracranial hemorrhage (5), and the fastMRI dataset from New York University and Facebook AI Research containing 10,000 knee MRI studies (6). Each of these datasets, other than the fastMRI dataset, are accompanied by labels (indicators of a particular disease or imaging

Acad Radiol 2020; 27:106–112

From the Australian Institute for Machine Learning, North Terrace, Adelaide, Australia (L.O.-R.); School of Public Health, University of Adelaide, North Terrace, Adelaide 5000, Australia (L.O.-R.); Royal Adelaide Hospital, North Terrace, Adelaide, Australia (L.O.-R.). Received January 16, 2019; revised October 3, 2019; accepted October 14, 2019. This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors. **Address correspondence to:** L. O.-R. e-mail: luke.oakden-rayner@adelaide.edu.au

https://doi.org/10.1016/j.acra.2019.10.006

finding within each study) that are intended to inform the training of clinically useful AI systems.

These public datasets have generated an enormous level of interest in the medical image analysis community. Two hundred and sixty teams registered for the RSNA Bone Age challenge, and over 1400 teams registered for the RSNA pneumonia detection challenge. Similarly, dozens of teams have published results on the CXR14 and MURA datasets.

This democratization of access to large-scale medical data has undoubtedly been of benefit to the medical image analysis community; however it is important to understand the specific challenges presented by these public datasets.

The root of the problem with public datasets is that the development processes (data gathering, cleaning, and labeling) are disconnected from the usage of the data. This means that the end-user of the data may not understand the nuances of the development processes, including many subtle design decisions that are not always well communicated in the published reports on the datasets. This problem is compounded by the highly opaque nature of medical images to nonexperts (i.e., nonradiologists). Unlike with datasets of ordered rows and columns of numbers, where the relationships between input variables and labels can be analyzed by anyone practiced in the skills of data science, the connection between a medical image and its label (for example, a diagnosis) requires domain understanding.

This disconnect between dataset development and usage can cause a variety of problems; (1) the accuracy of the labels can be overestimated by users, particularly when the weaknesses of the label generation procedures are poorly explained, (2) the presence of unlabeled visual subsets (subgroups of images that *look* different than the majority of images in the label class) can significantly alter the usefulness of the labels in training AI systems, and (3) the clinical meaning of the labels themselves can be obscure. Making matters worse, if AI systems are then tested on data generated with the same procedures (i.e., on test data drawn from the same dataset), then these problems may occur silently; the results of testing can look good because the models can learn to reproduce the flawed labels from the training data, but the actual clinical performance of these systems will be poor.

Each of these problems can only be overcome by the direct application of medical knowledge; an expert must appreciate the presence of subsets, review the quality of the labels, and comprehend the logic of the label schema (the rules that govern what each label means, and their relationships to each other). Only after this evaluation is it possible to determine the value of the dataset for building medical AI systems.

In this work we explore two large public datasets to demonstrate the importance of this review process, assessing the accuracy of the provided labels, as well as identifying other issues that may limit the utility of these datasets. In doing so, we stress the importance of expert visual analysis as a form of quality control when building and using these large-scale datasets, and present recommendations for teams planning to release public datasets in the future.

METHODS

Datasets

CXR14

The CXR14 dataset is a large-scale dataset for pathology detection in chest radiographs. This dataset was released in 2017 and updated later the same year, containing 112,120 frontal chest films from 30,805 unique patients. The dataset is drawn from a single tertiary medical center (the NIH Clinical Center) and appears to include films from multiple clinical settings, including intensive care unit (ICU) and non-ICU patients.

The images had a resolution of 3000 × 2000 pixels, and were in the DICOM format (which stores grayscale pixels with around 3000−4000 gray levels). These were downsampled into PNG images with a resolution of 1024 × 1024 pixels and 255 gray levels, an absolute reduction in complexity of around 99%.

The dataset was labeled using natural language processing on the original (clinical) free-text reports, a process that involved matching keywords related to various forms of pathology, and identifying negations (sentences that exclude certain findings and pathologies).

The dataset was initially labeled with eight different classes, however this was expanded to 14 classes later in 2017. These classes were: atelectasis, consolidation, infiltration, pneumonia, cardiomegaly, pneumothorax, fibrosis, pleural effusion, mass, nodule, pleural thickening, oedema, hiatus hernia, emphysema, and a normal (no finding) class. As the dataset was collected from a clinical archive, these image findings occur roughly at clinical prevalence, ranging from less than 0.5% (hernia) up to almost 10% (infiltration). The "normal" or "no finding" class makes up around 75% of the total images, or roughly 84,000 studies.

Notably, there are many patients with multiple X-rays. Almost half of all patients (13,302) had more than one study, together accounting for 84% of the data. Further exploration of the extent of multiple studies is provided in Table 1, revealing that the number of patients with numerous repeat images account for a surprisingly large proportion of the dataset. In many of these cases, these will reflect ICU patients who have repeat imaging daily, where the images change very little across the entire series, significantly reducing the diversity of the dataset.

MURA

The MURA dataset is a large dataset for abnormality detection in upper limb musculoskeletal radiographs. Released in 2018, the dataset contains 40,561 images from 14,863 studies, obtained from a single tertiary medical center (Stanford Hospital). The dataset includes seven standard upper limb study types, with studies of the fingers (2110 studies), hands (2185 studies), wrists (3697 studies), forearms (1010 studies), elbows (1912 studies), humeri (727 studies), and shoulders (3015 studies).

The dataset was labeled at the time of clinical interpretation by board-certified radiologists, each providing a label of

TABLE 1. The Prevalence of Patients with Multiple Studies in the CXR14 Dataset

Number of Studies per Patient	Number of Patients (Percentage of Patients in CXR14)	Total Number of Studies (Percentage of Studies in CXR14)
>1	13,302 (43%)	94,617 (84%)
>5	4821 (16%)	70,081 (63%)
>10	2225 (7%)	50,468 (45%)
>50	151 (0.5%)	10,812 (10%)
>100	18 (<0.001%)	2310 (2%)

"normal" or "abnormal" at the time of performing their usual report. No additional pathology specific labels were produced.

It was not specified in the MURA paper what constituted an "abnormal" finding, however an analysis of the label composition was performed by the authors; the text reports of 100 cases labeled abnormal were reviewed and the specific abnormalities identified were noted. This analysis revealed, of 100 abnormal cases, there were 53 studies with fractures, 48 studies with implanted hardware (such as joint replacements), 35 studies with degenerative joint disease, and 29 studies with other abnormalities such as lesions and subluxations.

The image in MURA was also downsampled, from DICOM images with a resolution of 1500 × 2000 pixels and around 3000−4000 gray levels, to PNG images with a resolution of 512 × 200−400 pixels, and 255 gray levels. This reflects an absolute reduction in image complexity of over 95%.

Multiple views were performed for most cases, with over 92% of cases having more than one image. There were relatively few repeat studies however, with only around 4% of patients having two or more studies.

Visual Inspection

The accuracy of the datasets was assessed by visual inspection performed by LOR, a board-certified radiologist. Each dataset was relabeled by LOR according to the findings of an initial exploratory assessment. Visual review of a randomly generated subset of around 100 images from each class (i.e., pneumonia, cardiomegaly, etc. in CXR14, and normal, abnormal in MURA) was performed. This exploration was to understand the images and labels, and to identify any common problems with the label schemata, with the findings used to inform the relabeling process. The images used for exploratory analysis were separate from the images that were used for relabeling.

CXR14

In the CXR14 dataset, large-scale relabeling at the original prevalence was not achievable due to the rarity of many of the labels present in the original dataset and the ambiguity of many of the label classes. For example the CXR14 label schema considers pneumonia, consolidation, and infiltration as distinct processes, but clinically it is rarely possible to distinguish these processes, at least without clinical information. As such, any labels created for these categories would likely be idiosyncratic and unfairly reflect upon the accuracy of the original labels.

To overcome these issues, an enriched, randomly selected subset of 50 cases per class were reviewed, for a total of 700 cases. Rather than attempting to relabel each case with the 14 possible classes, each case was reviewed purely for the presence of the label(s) it had been given; for example, a case from the cardiomegaly subset was relabeled "cardiomegaly" or "not cardiomegaly".

As each label class is not explicitly defined in the original paper, a permissive labeling rule was applied. In general, if a case could plausibly reflect the finding in the original label, it was considered positive for that finding, as long as that finding was visible to the eye of faith (i.e., an ill-defined basal opacity could be positive for pleural effusion or consolidation, but not for a mass). This is in comparison to labeling rules such as "label only findings that you would report in clinical practice" or "label only findings that you are certain of," both of which rules are much stricter than the rule applied. As such, the labels should diverge as little as possible from the original CXR14 labels, while still reflecting the visual appearance.

Notably this label rule was more permissive than the rules used in previous analysis of this dataset (7).

Because this approach was very permissive, a second rule which was closer to normal clinical practice was also applied for relabeling. The finer details of both of these rules are provided in the Supplement.

MURA

In MURA, relabeling of 714 randomly selected cases was performed. This was achievable at the original prevalence of the dataset, as the ratio of abnormal to normal cases was around 45:55.

Pathology specific labeling was undertaken, with labels produced to identify cases containing fractures, implanted hardware, degenerative joint disease, bone tumors, and a class containing miscellaneous pathologies (such as osteopaenia, subluxations, and ligamentous injuries). These labels were produced without knowing the original MURA label (normal or abnormal) for each study.

The labeling rule in this case was less permissive, as these labels were part of a larger effort to relabel the MURA dataset with clinically accurate labels. As such, cases were labeled to the best accuracy of the radiologist.

Analysis

The CXR14 data labels were assessed by calculating their positive predictive value (PPV), using the expert visual labels as the gold-standard. The PPV was presented here because the cases were only assessed for the presence or absence of their associated labels findings. As explained earlier, the

dataset could not be efficiently relabeled wholesale for a variety of reasons, which meant that the negative predictive value could not be determined. As such it should not be assumed that these results reflect a comprehensive assessment of the dataset, but instead simply provide evidence towards the quality of the labels.

The MURA data labels were assessed using the expert visual labels as the ground truth, with the sensitivity and specificity calculated. In this dataset, comprehensive relabeling was possible, which allowed for assessment of both positive and negative cases.

In both datasets, subgroup analysis was performed. In the MURA dataset, the subgroups included the body region imaged and the specific pathology groups labeled by visual review. In the CXR14 dataset, the specific subgroups beyond just the label classes were identified during exploratory analysis of the images.

RESULTS

CXR14

Fifty cases from each of the 15 class groups were assessed by LOR. The results for the visual assessment of the CXR14 dataset are provided in Table 2. Even with the use of permissive labeling rules, the PPV determined by visual assessment of the images is below the estimated PPV presented in Wang et al. in all classes.

Exploratory visual analysis revealed two striking examples of visual stratification. The first is in the pneumothorax class, where 80% of the positive cases have chest drains. In these examples, there were often no other features of pneumothorax (i.e., the lung did not appear collapsed, likely reflecting a successfully treated pneumothorax). While the overall PPV was quite high (90%), of the cases without chest drains the PPV was lower, at 60%.

The second example of visual stratification was related to the emphysema class. The majority of cases (86%) had subcutaneous emphysema rather than pulmonary emphysema. This is almost certainly a specific failure of the original labeling method, where these keywords were not successfully disambiguated. This resulted in a very low PPV for the emphysema labels.

TABLE 2. Visual Assessment of the CXR14 Dataset Labels, Using Both Permissive and "Clinical-Style" Relabelling Rules as the Ground-Truth

	PPV (Visual, Permissive)	PPV (Visual, Clinical)	PPV (Text Mining, From Wang et al.)††
Consolidation	80%	66%	-
Atelectasis	80%	50%	99%
Infiltration	66%	36%	74%
Pneumonia	60%	50%	66%
Oedema	76%	40%	-
Nodule	76%	64%	96%
Mass	64%	46%	75%
Pneumothorax*	90% (60%)	90% (60%)	90%
Pleural effusion	74%	70%	93%
Pleural thickening	84%	52%	-
Emphysema	14%	10%	-
Cardiomegaly	70%	52%	100%
Fibrosis	46%	26%	-
Hernia†	94%	78%	-
Normal	76%	62%	87%

* The pneumothorax class was stratified, with the majority of images containing chest drains. The PPV of the subset of cases without chest drains is given in parentheses.

† The hernia class was almost always correctly labeled, but on exploratory analysis there were many examples of false negatives that were not captured in the PPV value.

†† The text mining PPV reported by Wang et al. was scored against the Open-i dataset (9), rather than using data from the CXR14 cohort. The labels with no text mining PPV did not appear in the Open-i dataset.

MURA

The imaging characteristics of the test subset, as well as the entire MURA dataset, are provided in Table 3. The distribution of cases within these groups was similar.

Taking the expert review of the images as the ground-truth, the sensitivity (true positive rate) and specificity (true negative rate) of the MURA labels is presented in Table 4, both overall and by region.

The class specific (i.e., per pathology) sensitivity of the MURA labels is presented in Table 5, using the expert class labels as the ground truth. The specificity of the MURA labels is also presented, but the values are inflated by the low per-class prevalence of the conditions relative to the number of normal studies.

There was poor identification of degenerative joint disease by the MURA labels (sensitivity = 60%) compared to fractures and hardware (sensitivity = 92% and 85%, respectively), which was also reflected in lower sensitivity for identifying pathology

TABLE 3. The Imaging Characteristics of the Relabelled Test Subset, and the MURA Dataset Overall

	No. in Test Subset (% of the Subset)	Percentage of the Test Subset Labelled Abnormal	No. in MURA Dataset (% of the MURA Dataset)	Percentage of MURA Labelled Abnormal
All images	714 (100%)	43%	14656 (100%)	39%
Finger	126 (18%)	44%	2110 (14%)	35%
Hand	131 (18%)	44%	2185 (15%)	27%
Wrist	182 (25%)	36%	3697 (25%)	38%
Forearm	45 (6%)	28%	1010 (7%)	35%
Elbow	70 (10%)	33%	1912 (13%)	38%
Humerus	19 (3%)	42%	727 (5%)	47%
Shoulder	141 (20%)	52%	3015 (21%)	52%

TABLE 4. The Sensitivity and Specificity of the MURA Labels, Overall and by Region. Specific Subgroups Where the Labels Underperform Compared to the Average Performance Across the Dataset are Highlighted in Bold

	Sensitivity	Specificity
All images	80%	75%
Finger	**72%**	82%
Hand	**79%**	80%
Wrist	**63%**	89%
Forearm	**56%**	90%
Elbow	94%	81%
Humerus	100%	92%
Shoulder	82%	**64%**

TABLE 5. The Sensitivity and Specificity of the MURA Labels by Pathology Type. The Specificity Values Appear High Partially Due to the Low Per-Class Prevalence of the Findings

	Sensitivity	Specificity
Fracture	92%	98%
Hardware	85%	98%
Degenerative disease	60%	82%
Other	82%	97%

in the regions typically affected by joint disease (i.e., in the wrist and the forearm, but not in the elbow or humerus).

To further explore this correlation, a linear regression model was created to quantify the relationship between the prevalence of joint disease by region, and the sensitivity for the detection of abnormal findings by region. These results are presented in Figure 1, demonstrating a clear relationship (the coefficient is −0.88). In other words, the more joint disease occurring in the region, the worse the labels are for identifying abnormal cases.

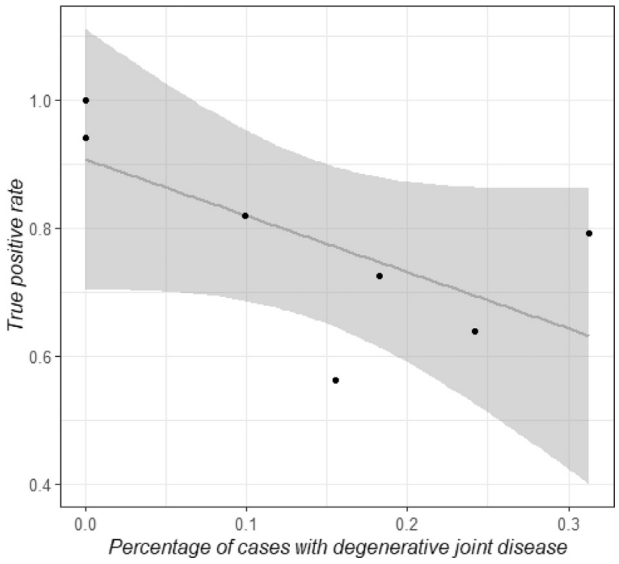


Figure 1. The relationship between sensitivity (true positive rate) and the prevalence of degenerative joint disease in the images from specific regions (i.e., hand, wrist, etc.).

During this analysis, it was noted that there was an unexpected number of false positive results in the shoulder class; that is, cases that were labeled abnormal in MURA but were considered normal on visual review. Visual exploration of these cases revealed no clear patterns. There were several missed diagnoses amongst the visually reviewed cases (several subtle fractures and two lytic bone lesions), but the majority of the false positive cases revealed no identifiable pathology.

DISCUSSION

The two datasets explored were of variable quality. The PPV of the labels in the CXR14 dataset were typically quite low, even allowing for differences in reporting style and interobserver variability. By contrast, the MURA labels were of much higher accuracy, other than in the subset of patients with features of degenerative joint disease.

In both datasets, the errors in the labels appear directly related to the weaknesses of the respective labeling methods.

In the CXR14 data, the use of natural language processing on the reports is problematic because even if the process of label extraction is flawless, the reports themselves are often incomplete descriptions of the images. This hypothesis is supported by the large gap between the findings in Wang et al., which show that their labels are accurate reflections of the reports, and the visual appearance of the images. This discrepancy is understandable from a clinical perspective, as radiology reports are not simply an enumeration of image findings. Many findings are not included in radiology reports either because they are already known (for example, the classic report that only states "no change compared to previous") or because the radiologist determined that the finding was not relevant to the referring clinician. This is a major concern because it suggests that label harvesting with natural language processing may never be able to accurately reflect the image findings on the films, and thus may never be able to produce high quality labels for training image analysis systems.

The CXR14 dataset also included examples of label disambiguation failure, with the majority of cases labeled "emphysema" actually showing evidence of subcutaneous emphysema, and of label schema failure, where the labels did not account for the clinically important stratification in the pneumothorax class. The majority of pneumothorax cases were already treated with chest tubes and often did not show radiographic evidence of ectopic pleural gas.

This latter point is important not only because untreated pneumothoraces are more clinically important to identify, but also because these labels are intended to train image analysis systems. While it is technically true that a patient with a chest tube in "has a pneumothorax," if the majority of cases do not show any of the visual features associated with this pathology, the usefulness of the labels is highly suspect. What can we reasonably expect models to learn from these labels, other than the appearance of chest tubes?

The label schema of the CXR14 data also suffered from significant ambiguity, particularly related to the various labels for airspace opacities, to the point that it became almost impossible to design an acceptable way to relabel these classes. This was noted on review of the CXR14 labels themselves, as it was highly unclear why a case would be labeled as one class but not another. No common patterns were identified with exploratory analysis.

The MURA labeling process was much more robust, because each radiologist at the point of care was asked to label each case as normal or abnormal. The issue with this labeling strategy arises because the definition of normal and abnormal cases was left up to each radiologist. Anecdotally, the Stanford team has suggested in private communication that many of the radiologists interpreted this to mean "normal *for age*". This would be in keeping with the finding that degenerative joint disease in particular was under-reported in the labeling process, as presumably many radiologists may have decided that minor degenerative disease was within the expected range of normal for older patients. As the age of the patients is not provided with the dataset, this hypothesis cannot be explored further.

These weaknesses do not necessarily detract from the general usefulness of these datasets, but they do need to be understood if the models trained on the data are to function as expected. For example, a model trained on the MURA data should not be expected to detect hand or wrist osteoarthritis to any degree of accuracy. With this in mind, the biggest limitation of these datasets is not their label quality, but their documentation.

The supporting documents for these datasets do not adequately discuss these issues. In fact, the ChestXray14 paper and dataset FAQ explicitly state that "the text-mined disease labels are expected to have accuracy >90%." Similarly, the MURA dataset paper presents an exploration of 100 abnormal cases, and states that the abnormalities include fractures, hardware, degenerative changes, and other miscellaneous findings. This gives the impression that the labels do cover degenerative disease, but is in fact an artefact of the process of only looking at the abnormal cases. If the team had reviewed the cases labeled normal, they would likely have discovered the presence of many cases of degenerative joint disease in this group as well.

In both circumstances, the original documentation is misleading. This raises an important question; "who is responsible for ensuring the quality of the data in public datasets?" The effort required to explore a dataset of this size is not negligible, and we may fear that expecting this level of analysis of teams who intend to release public datasets may dissuade them from producing these important resources.

It is also true that many end-users of this data are teams of computer scientists and engineers, who may not have easy access to the medical expertise required to understand the nuances of the data. Even worse, due to this lack of expertise, they may not even realize that the data could be flawed in the first place, particularly if they rely on test sets drawn from the same data.

On balance, the effort required to manually inspect a small subset of a dataset is fairly low compared to the effort required to *build* such a dataset, if performed a single time in a centralized manner (i.e., at the time of building the dataset). The team that builds the dataset is ideally suited to performing this analysis because they already understand the data generating process (for example, the MURA team already had anecdotal knowledge of how the labeling rule was being applied) and already have access to the medical expertise required for this assessment.

One way to partially mitigate the problems that users of the data may face is to produce a smaller second dataset purely for testing models trained on the original data, using a less flawed method, ideally involving expert visual review of cases. The MURA team has done this, using the majority vote of three board-certified radiologists to produce the labels for 207 randomly selected cases. Unfortunately, no analysis of the quality of the original labels for these 207 cases was presented in the MURA paper, nor was any subset analysis done on these test labels.

The CXR14 team did not provide a manually labeled test set. Independently, a team that published results on this dataset produced their own visual labels (8), and showed that the original labels significantly underperformed compared to other radiologists tested on the new labels (F1 score of 0.288 vs radiologist F1 scores of 0.35−0.44). Unlike the MURA test set, these CXR14 test labels are not publically available.

In both cases, the labeling rules used to produce these test sets were not explicitly stated.

While the use of a visually accurate set of labels for the test set does not solve many of the issues of the primary dataset, it does protect against the risk that the models will fail silently; the insidious risk that the model can reproduce the flawed labels but appear to be performing well because the test set is equally as flawed.

There are a number of limitations to this analysis that should be acknowledged. First of all, the labels produced by LOR are not 100% accurate. There will always be a significant amount of inter-observer variability, particularly when labels are ambiguous. This was compounded by the reduced image quality in each dataset. In particular, the reduction in the number of gray levels meant that many dense parts of the images became completely obscured, as if only a single window setting was available for review. For example it was regularly impossible to identify any retrocardiac pathology in the CXR14 dataset, because the heart appeared purely white.

These limitations were mitigated to some extent by being as permissive as possible when relabeling the CXR14 data, erring on the side of agreeing with the original labels.

It is also true that some of the labels in the CXR14 and MURA datasets are informed by information not available to the reviewer. This is probably particularly true in the case of airspace opacities, where a label of pneumonia or consolidation

may be more likely to be used in a patient with a fever. However, in the context of producing labels for image analysis systems, it may actually be the case that a blind review of the images themselves is more worthwhile, as the presence or absence of image features alone is all that the models will be able to learn.

CONCLUSION

The disconnect between the dataset development and the usage of that data can lead to a variety of major problems in public datasets. The accuracy, meaning, and clinical relevance of the labels can be significantly impacted, particularly if the dataset development is not explained in detail and the labels produced are not thoroughly checked.

These problems can be mitigated by the application of expert visual review of the label classes, and by thorough documentation of the development process, strengths, and weaknesses of the dataset. This documentation should include an analysis of the visual accuracy of the labels, as well as the identification of any clinically relevant subsets within each class. Ideally, this analysis and documentation will be part of the original release of the data, completed by the team producing the data to prevent duplication of these efforts, and a separate test set with visually accurate labels will be released alongside any large-scale public dataset.

REFERENCES

1. Sun C, Shrivastava A, Singh S, Gupta A. Revisiting unreasonable effectiveness of data in deep learning era. In: Computer vision (ICCV), 2017 IEEE international conference on; 2017, IEEE; 2017:843–852.
2. Wang X, Peng Y, Lu L, et al. Chestx-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. In: 2017 IEEE conference on computer vision and pattern recognition (CVPR); 2017, IEEE; 2017:3462–3471.
3. Rajpurkar P, Irvin J, Bagul A, et al. Mura dataset: towards radiologist-level abnormality detection in musculoskeletal radiographs. *arXiv preprint arXiv:1712069572*017.
4. Halabi SS, Prevedello LM, Kalpathy-Cramer J, et al. The RSNA pediatric bone age machine learning challenge. Radiology 2018; 290(2):498–503.
5. Chilamkurthy S, Ghosh R, Tanamala S, et al. Deep learning algorithms for detection of critical findings in head CT scans: a retrospective study. Lancet 2018; 392(10162):2388–2396.
6. Zbontar J, Knoll F, Sriram A, et al. fastmri: an open dataset and benchmarks for accelerated mri. *arXiv preprint arXiv:1811088392*018.
7. Oakden-Rayner LExploring the ChestXray14 dataset: problems. 2018.
8. Rajpurkar P, Irvin J, Zhu K, et al. Chexnet: radiologist-level pneumonia detection on chest x-rays with deep learning. *arXiv preprint arXiv:1711052252*017.
9. Demner-Fushman D, Kohli MD, Rosenman MB, et al. Preparing a collection of radiology examinations for distribution and retrieval. J Am Med Inform Assoc 2015; 23(2):304–310.

SUPPLEMENTARY MATERIALS

Supplementary material associated with this article can be found in the online version at doi:10.1016/j.acra.2019.10.006.

1.2 Underperforming data subsets

While the accuracy of the labels is well-known to be important, the variability *within* label groups is less often recognised as a source of the implementation gap. Labelling schemas are designed based on a number of factors, and while exhaustive labeling of all features in an image is desirable this process is costly and time consuming, and dataset developers are limited by the availability of labeling resources (i.e., expert clinicians). Limiting labelling to coarse superclasses which describe broad constellations of image features is a widely employed strategy, for example labelling "lung masses" on a chest x-ray rather than specifically labelling subtypes or even image features (such as "spiculated" vs "rounded" masses, or "solid" vs "cavitating" lesions).

When overly broad class labels are used there is a high chance to produce an implementation gap; if smaller subsets are associated with worse performance then the use of such datasets can produce poor outcomes, especially if, as is often the case in medical imaging, serious or life-threatening conditions tend to be less common than more benign variants. In this context, not only does testing fail to detect underperformance within the smaller subset (as the majority subclass obscures the minority in aggregate performance metrics), but medical AI models are naturally less likely to perform well on these subsets because the rare subset has fewer training examples, and therefore provides a weaker training signal to the model.

In "**Hidden stratification causes clinically meaningful failures in machine learning for medical imaging"** [31] I explore the role of unrecognised subsets in AI systems, including how the size of minority subclasses, the quality of subclass labels, and the visibility/subtleness of subclass features can result in "hidden" underperformance in important clinical tasks. I also identify several possible ways to discover subsets when developing datasets and testing AI models.

Hidden stratification causes clinically meaningful failures in machine learning for medical imaging

Publication status: Published in the *Proceedings of the ACM Conference on Health, Inference, and Learning, 2020.*

Contribution: 40%

Detailed description contribution: This inter-disciplinary work required equal input from myself and my co-first-author Dr Jared Dunnmon, who was a post-doctoral machine learning researcher at the time of the work. We each contributed an estimated 40% of the total work towards the publication, including in planning and study design, data collection and cleaning, experimentation, manuscript drafting and editing.

Senior authors Professors Chistopher Re and Gustavo Carneiro each were involved in planning and study design as well as manuscript editing, and each contributed an estimated 10% of the work towards the publication.

Certification from co-authors:

Christopher M. Ré.

Jared Dunnmon.

Declaration: This publication was part of the work undertaken during and for my HDR candidature.

Hidden Stratification Causes Clinically Meaningful Failures in Machine Learning for Medical Imaging

Luke Oakden-Rayner*
luke.oakden-rayner@adelaide.edu.au
Australian Institute for Machine Learning
University of Adelaide
Adelaide, Australia

Jared Dunnmon*
jdunnmon@cs.stanford.edu
Department of Computer Science
Stanford University
Stanford, California, USA

Gustavo Carneiro
gustavo.carneiro@adelaide.edu.au
Australian Institute for Machine Learning
University of Adelaide
Adelaide, Australia

Christopher Ré
chrismre@cs.stanford.edu
Department of Computer Science
Stanford University
Stanford, California, USA

ABSTRACT

Machine learning models for medical image analysis often suffer from poor performance on important subsets of a population that are not identified during training or testing. For example, overall performance of a cancer detection model may be high, but the model may still consistently miss a rare but aggressive cancer subtype. We refer to this problem as *hidden stratification*, and observe that it results from incompletely describing the meaningful variation in a dataset. While hidden stratification can substantially reduce the clinical efficacy of machine learning models, its effects remain difficult to measure. In this work, we assess the utility of several possible techniques for measuring hidden stratification effects, and characterize these effects both via synthetic experiments on the CIFAR-100 benchmark dataset and on multiple real-world medical imaging datasets. Using these measurement techniques, we find evidence that hidden stratification can occur in unidentified imaging subsets with low prevalence, low label quality, subtle distinguishing features, or spurious correlates, and that it can result in relative performance differences of over 20% on clinically important subsets. Finally, we discuss the clinical implications of our findings, and suggest that evaluation of hidden stratification should be a critical component of any machine learning deployment in medical imaging.

CCS CONCEPTS

• **Computing methodologies** → **Machine learning**.

KEYWORDS

hidden stratification, machine learning, convolutional neural networks

*Both authors contributed equally to this research.

ACM CHIL '20, April 2–4, 2020, Toronto, ON, Canada

ACM ISBN 978-1-4503-7046-2/20/04.
https://doi.org/10.1145/3368555.3384468

ACM Reference Format:
Luke Oakden-Rayner, Jared Dunnmon, Gustavo Carneiro, and Christopher Ré. 2020. Hidden Stratification Causes Clinically Meaningful Failures in Machine Learning for Medical Imaging. In *ACM Conference on Health, Inference, and Learning (ACM CHIL '20), April 2–4, 2020, Toronto, ON, Canada.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3368555.3384468

1 INTRODUCTION

Deep learning systems have shown remarkable promise in medical image analysis, often claiming performance rivaling that of human experts [13]. However, performance results reported in the literature may overstate the clinical utility and safety of these models. Specifically, it is well known that machine learning models often make mistakes that humans never would, despite having aggregate error rates comparable to or better than those of human experts. An example of this "inhuman" lack of common sense might include a high performance system that calls any canine in the snow a wolf, and one on grass a dog, regardless of appearance [31]. This property of machine learning models is likely to be of critical importance in medical practice, where specific types of errors can have serious clinical impacts.

Of particular concern is the fact that most medical machine learning models are built and tested using an incomplete set of possible labels—or *schema*—and that the training labels therefore only coarsely describe the meaningful variation within the population. Medical images contain dense visual information, and imaging diagnoses are usually identified by recognizing the combination of several different visual features or patterns. This means that any given pathology or variant defined as a "class" for machine learning purposes is often comprised of several visually and clinically distinct subsets; a "lung cancer" label, for example, would contain both solid and subsolid tumors, as well as central and peripheral neoplasms. We call this phenomenon *hidden stratification*, meaning that the data contains unrecognized subsets of cases which may affect model training, measured model performance, and most importantly the clinical outcomes related to the use of a medical image analysis system.

Worryingly, when these subsets are not labelled, even performance measurements on a held-out test set may be falsely reassuring. This is because the aggregate performance measures such

as sensitivity (i.e. recall) or ROC-AUC can be dominated by larger subsets, obscuring the fact that there may be an unidentified subset of cases within which performance is poor. Given the rough medical truism that serious diseases are less common than mild diseases, it is even likely that underperformance in minority subsets could lead to disproportionate harm to patients.

In this article, we demonstrate that hidden stratification is a fundamental technical problem that has important implications for medical imaging analysis on multiple real-world datasets, and explore several possible techniques for measuring its effects. We first define three distinct approaches to measuring hidden stratification effects—schema completion, error auditing, and algorithmic measurement—and detail their relative strengths and weaknesses. We next apply schema completion to illustrate that hidden stratification is present in standard computer vision models trained on the CIFAR-100 benchmark dataset, and leverage this well-characterized dataset to empirically explore several possible causes of hidden stratification. We extend our analysis to medical imaging datasets by using a combination of schema completion, error auditing, and algorithmic measurement to show not only that hidden stratification can result in performance differences of up to 20% on clinically important subsets, but also that simple unsupervised learning approaches can help to identify these effects. Using these measurement techniques, we find evidence across multiple datasets that hidden stratification occurs on subsets characterized by a combination of low prevalence, poor label quality, subtle discriminative features, and spurious correlates.

Our results also suggest that more research is necessary on methods for mitigating hidden stratification. Though we show that approaches that require manual application of human expertise—such as schema completion and error auditing—have potential in practice, widespread use of such techniques is likely to be limited by cost. Algorithmic measurement approaches, on the other hand, require more limited human intervention, but are of variable performance and cannot yet guarantee the detection of important subsets. We examine the clinical implications of these findings, and argue that despite the associated challenges, measurement and reporting of hidden stratification effects should become a critical component of machine learning deployments in medicine.

Our paper is organized as follows: Section 2 contextualizes the hidden stratification problem with respect to related work, Section 3 defines three possible methods by which hidden stratification can be measured, Section 4 presents the results of experiments that apply these measurement techniques to multiple relevant datasets, and Section 5 discusses both the clinical implications of our findings and the limitations of our study.

2 RELATED WORK

Problems similar to hidden stratification have been observed or postulated in many domains, including traditional computer vision [30], fine-grained image recognition [39], genomics [7], and epidemiology (often termed "spectrum effects") [23]. The difficulty of the hidden stratification problem fundamentally relates to the challenge of obtaining labelled training data. Were fine-grained labels available for every important variant that could be distinguished via a given data modality, discriminative model performance on important subsets could be improved by training and evaluating models using this information. Thus, typical approaches to observed stratification and dataset imbalance in medical machine learning often center on gathering more data on underperforming subsets, either via additional labelling, selective data augmentation, or oversampling [22]. However, the cost of manual labelling is often prohibitive, appropriate augmentation transforms can be difficult to define, and oversampling an underperforming subset can cause degradation on others [4, 14, 29, 41]. As a result, medical imagery analysts have commonly begun either to use semi-automated labelling techniques [10, 14, 18, 35] or to apply human expertise to produce a narrow or incomplete set of visual labels [26] rather than exhaustively labelling all possible findings and variations. Both of these approaches can yield reduced accuracy on important subsets [24]. Techniques that reliably increase performance on critical imaging subsets without degrading performance on others have yet to be demonstrated.

Methods that directly address hidden stratification, where the subclasses are obscure, have not been commonly explored in medical imaging analysis. However, it is clear from the recent literature that this issue has been widely (but not universally) recognized. The most common approach for measuring hidden stratification is by evaluating model performance on specific subsets. Gulshan et al. [16], for instance, present variations in retinopathy detection performance on subsets with images obtained in different locations, with differing levels of disease severity, and with different degrees of pupil dilation. In several cases, their models perform differently on these subsets in a manner that could be clinically impactful. Chilamkurthy et al. [9] present a subset analysis for different diagnostic categories of intracranial hemorrhage (e.g. subdural vs. subarachnoid) when designing a deep learning model for abnormality detection on head CT, but do not analyze differences in performance related to bleed size, location, or the acuity of the bleed. Their work does, however, evaluate the performance of models on cases with multiple findings, and observe substantial variation in model performance within different strata; for instance, subarachnoid bleed detection performance appears to degrade substantially in the presence of an epidural hemorrhage. Wang et al. [34] perform an excellent subset analysis of a colonoscopy polyp detector, with comparative performance analysis presented by polyp size, location, shape, and underlying pathology (e.g. adenoma versus hyperplastic). Similarly, Dunnmon et al. [11] report the performance of their chest radiograph triage system by pathology subtype, finding that models trained on binary triage labels achieved substantially lower performance on fracture than on other diseases. Non-causal confounding features such as healthcare process quantities can also contribute substantially to high model performance on data subsets heavily associated with these confounding variables [1, 2, 36, 41].

Instead of analyzing subsets defined *a priori*, Mahajan et al. [21] describe algorithmic audits, where detailed examinations of model errors can lead to model improvements. Several recent studies perform error audits, where specific failure modes such as small volume cancers, disease mimics, and treatment-related features are observed [6, 34]; such analyses may be helpful in identifying error modes via human review, but do not characterize the full space of subset performance [33]. There has been particular interest in

formalizing algorithmic audit methods recently [25], although these initiatives have yet to be tailored to the medical setting.

Of course, there also exist multiple studies that do not directly address the effects of hidden stratification [3, 17]. Esteva et al. [12] is particularly notable, as this dataset is labelled for more than 2,000 diagnostic subclasses but the results presented only consider "top-level" diagnostic categories. Analysis of these effects would improve the community's ability to assess the real-world clinical utility of these models.

3 METHODS FOR MEASURING HIDDEN STRATIFICATION

We explicitly define and evaluate three possible approaches to measure the clinical risk of hidden stratification: (1) exhaustive prospective human labeling of the data, called *schema completion*, (2) retrospective human analysis of model predictions, called *error auditing*, and (3) automated *algorithmic measurement* methods to detect hidden strata. Each of these methods is applied to the test dataset, allowing for analysis and reporting (e.g., for regulatory processes) of subclass (i.e. subset) performance.

Schema Completion: In schema completion, the schema author prospectively prescribes a more complete set of subclasses that need to be labeled, and provides these labels on test data. Schema completion has many advantages, such as the ability to prospectively arrive at consensus on subclass definitions (e.g. a professional body could produce standards describing reporting expectations) to both enable accurate reporting and guide model development. However, schema completion is fundamentally limited by the understanding of the schema author; if important subclasses are omitted, schema completion does not protect against important clinical failures. Further, it can be time consuming (or practically impossible!) to exhaustively label all possible subclasses, which in a clinical setting might include subsets of varying diagnostic, demographic, clinical, and descriptive characteristics. Finally, a variety of factors including the visual artifacts of new treatments and previously unseen pathologies can render existing schema obsolete at any time.

Error Auditing: In error auditing, the auditor examines model outputs for unexpected regularities, for example a difference in the distribution of a recognizable subclass in the correct and incorrect model prediction groups. Advantages of error auditing include that it is not limited by predefined expectations of schema authors, and that the space of subclasses considered is informed by model function. Rather than having to enumerate every possible subset, only subsets observed to be concerning are measured. While more labor-efficient than schema completion, error auditing is critically dependent on the ability of the auditor to visually recognize differences in the distribution of model outputs. It is therefore more likely that the non-exhaustive nature of audit could limit certainty that all important strata were analyzed. Of particular concern is the ability of error auditing to identify low-prevalence, high discordance subsets that may rarely occur but are clinically salient.

Algorithmic Measurement: In algorithmic measurement approaches, the algorithm developer designs a method to search for subclasses automatically. In most cases, such algorithms will be unsupervised methods such as clustering. If any identified group (e.g. a cluster) underperforms compared to the overall superclass, then this may indicate the presence of a clinically relevant subclass. Clearly, the use of algorithmic approaches still requires human review in a manner that is similar to error auditing, but is less dependent on the specific human auditor to initially identify the stratification. While algorithmic approaches to measurement can reduce burden on human analysts and take advantage of learned encodings to identify subsets, their efficacy is limited by the separability of important subsets in the feature space analyzed.

4 EXPERIMENTS

In our experiments, we empirically measure the effect of hidden stratification using each of these approaches, and evaluate the characteristics of subsets on which these effects are important. Drawing from the existing machine learning literature, we hypothesize that there are several subset characteristics that contribute to degraded model performance in medical imaging applications: (1) low subset prevalence, (2) reduced label accuracy within the subset, (3) subtle discriminative features, and (4) spurious correlations [33]. These factors can be understood quite simply: if the subset has few examples or the training signal is noisy, then the expected performance will be reduced. Similarly, if one subset is characterized by features that are harder to learn, usual training procedures result in models that perform well on the "easy" subset. Finally, if one subset contains a feature that is correlated with the true label, but not causal, models often perform poorly on the subset without the spurious correlate.

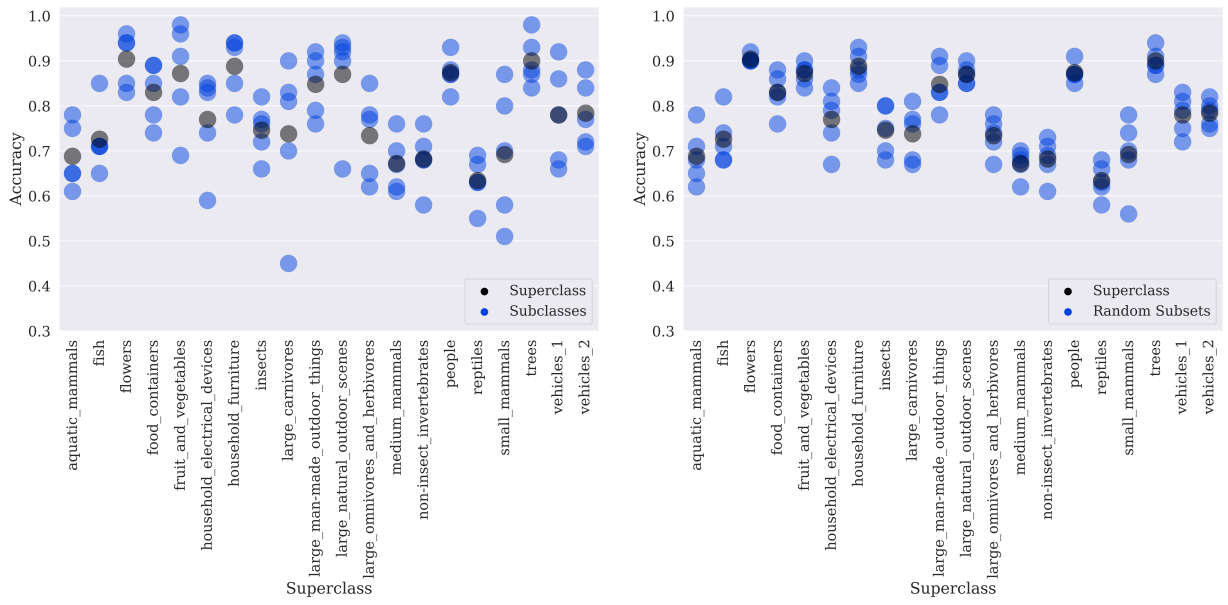
To demonstrate the technical concept of hidden stratification in a well-characterized setting, we first use schema completion to demonstrate substantial hidden stratification effects in the CIFAR-100 benchmark dataset, and confirm that low subset prevalence and reduced subset label accuracy can reduce model performance on subsets of interest. We then use this same schema completion technique to evaluate clinically important hidden stratification effects in radiograph datasets describing hip fracture (which contains subsets with low prevalence and subtle discriminative features) and musculoskeletal extremity abnormalities (which contains subsets with poor label quality and subtle discriminative features). Each of these datasets has been annotated a priori with labels for important subclasses, and is thus amenable to schema completion. We then demonstrate how error auditing can be used to identify hidden stratification in a large public chest radiograph dataset that contains a spurious correlate. Finally, we show that a simple unsupervised clustering algorithm can provide value by separating the well-performing and poorly-performing subsets identified by our previous analysis.

Code describing these experiments is available at `www.github.com/HazyResearch/hidden-stratification-mi`.

4.1 Schema Completion

We first use schema completion to measure the effects of hidden stratification on CIFAR-100 [19], Adelaide Hip Fracture [15], and MURA [26] datasets. When feasible, even partial schema completion can be useful for assessing hidden stratification.

CIFAR-100: The benchmark CIFAR-100 dataset from computer vision represents an excellent testbed on which to demonstrate the



(a) True (Semantic) Subclasses **(b) Random Subclasses**

Figure 1: Performance of a ResNeXt-29, 8x64d on CIFAR-100 superclasses by (a) true (semantic) CIFAR-100 subclass and (b) random CIFAR-100 subclasses. Random subclasses were assigned by randomly permuting the subclass label assignments within each superclass. Most superclasses contain true subclasses where performance is far lower than that on the aggregate superclass. Intra-subclass performance variance on random subclasses is on average 66% lower than on the true (semantic) subclasses, indicating that the stratification observed in practice is substantially higher than would be expected from randomness alone.

effect of hidden stratification in a well-characterized environment [19]. The CIFAR-100 dataset consists of 60,000 images binned into 20 "superclasses," which each contain five distinct "subclasses." Each subclass is represented in the dataset with equal frequency. We hypothesize that by training models only on superclass labels, and assessing superclass performance within each subclass, we will commonly observe subclasses on which performance is substantially inferior to that of the overall superclass. We further expect that subclass performance will degrade if that subclass is subsampled or if noise is added to superclass labels for that subclass, simulating stratification with low subclass prevalence or reduced label accuracy. For the purposes of this experiment, we assume that the CIFAR-100 subclasses represent a reasonable attempt at schema completion, and measure superclass accuracy within each subclass.

Figure 1(a) presents the performance of a ResNeXt-29, 8x64d Convolutional Neural Network (CNN) trained on the 20 CIFAR-100 superclasses using the training schedule reported in [37] and the implementation provided by [38]. In each superclass, the five constituent subclasses exhibit substantial performance variation, and the worst-performing subclass can underperform the aggregate superclass by over 30 accuracy points. This same phenomenon in medical imaging would lead to massively different outcomes for different subsets of the population, be these demographically or pathologically determined. To confirm that these large differences in subclass performance do not result from random variation within each superclass, we randomly permute the subclass labels within each superclass and evaluate our model on these random subclasses. If random variation was the cause of the stratification observed in Fig. 1(a), we would expect the inter-subclass performance variance to remain unchanged in this experiment. Instead, we find that inter-subclass performance variance is reduced by an average of 66% across all superclasses when the subclasses are randomly rather than semantically assigned, indicating that the performance stratification observed in Fig. 1(a) cannot be attributed to random variation.

Table 1 (middle) shows classification results on randomly selected subclasses ("dolphin" and "mountain") when 75% of the examples in a subclass are dropped from the training set, simulating a subclass with reduced prevalence. While the overall marine mammals superclass performance drops by only 4 accuracy points when the dolphin subclass is subsampled, performance on the dolphin subclass drops by 14 points from 0.78 to 0.64. Similar trends are observed for the mountain subclass, where overall superclass performance drops by 5 accuracy points when the mountain subclass is subsampled, but performance on the mountain subclass itself drops by 19 points. Clearly, unmeasured subclass underrepresentation can lead to substantially worse performance on that subclass, even when superclass performance is only modestly affected.

We show a similar trend in Table 1 (right) when random noise is added to the labels of a given subclass by replacing the 25% of the

Subclass	Baseline Superclass	Baseline Subclass	Subsample Superclass	Subsample Subclass	Random Noise Superclass	Random Noise Subclass
Dolphin	0.69	0.78	0.65 (-4)	0.64 (-14)	0.67 (-2)	0.73 (-5)
Mountain	0.87	0.90	0.82 (-5)	0.71 (-19)	0.82 (-5)	0.73 (-17)

Table 1: Accuracy of a ResNeXt-29, 8x64d trained using the full CIFAR-100 dataset ("Baseline") and two synthetic experiments with altered datasets. ("Subsample") drops 75% of the dolphin and mountain subclasses from the training dataset, and ("Random Noise") assigns 25% of examples from these subclasses a random superclass label. Results reported are on superclass labels for the validation set. Numbers in parentheses are reductions in performance with respect to the baseline model for each experimental condition.

true superclass labels with a random incorrect label, simulating a subclass with reduced label accuracy. Performance on both dolphin and mountain subclasses drops substantially when label accuracy decreases; while overall superclass performance in each case drops by less than 5 points, subclass performance decreases by up to 17 points. Such stratification of label quality by pathology is highly likely to occur in medical datasets, where certain pathologies are easier to identify than others.

Adelaide Hip Fracture Schema completion also shows hidden stratification on a large, high quality pelvic x-ray dataset from the Royal Adelaide Hospital [15]. A DenseNet model previously trained on this dataset to identify hip fractures achieved extremely high performance (AUC = 0.994) [15]. We hypothesize that reduced subclass performance will occur even in models with high overall superclass performance, particularly in subclasses characterized by subtle visual features or low subclass prevalence. The distribution of the location and description subclasses is shown in Table 2, with subclass labels produced by a board-certified radiologist (LOR). We indeed find that sensitivity on both subtle fractures and low-prevalence cervical fractures is significantly lower ($p<0.01$) than that on the overall task. ROC curves for each of these subclasses and the overall superclass shown in Fig. 2(a) demonstrate that these differences in sensitivity would be expected across a variety of potential operating points. These results support the hypothesis that both subtle discriminative features and low prevalence can contribute to clinically relevant stratification.

Subclass	Prevalence (Count)	Sensitivity
Overall	1.00 (643)	0.981
Subcapital	0.26 (169)	0.987
Cervical	0.13 (81)	**0.911**
Pertrochanteric	0.50 (319)	0.997
Subtrochanteric	0.05 (29)	0.957
Subtle	0.06 (38)	**0.900**
Mildly Displaced	0.29 (185)	0.983
Moderately Displaced	0.30 (192)	1.000
Severely Displaced	0.36 (228)	0.996
Comminuted	0.26 (169)	1.000

Table 2: Superclass and subclass performance for hip fracture detection from frontal pelvic x-rays. Bolded subclasses show significantly worse performance ($p<0.01$) than that on the overall task.

MURA: We next use schema completion to demonstrate the effect of hidden stratification on the MURA musculoskeletal x-ray dataset developed by Rajpurkar et al. [26], which provides labels for a single class, identifying cases that are "normal" and "abnormal." These labels were produced by radiologists in the course of their normal work, and include visually distinct abnormalities such as fractures, implanted metal, bone tumors, and degenerative joint disease. These binary labels have been previously investigated and relabelled with subclass identifiers by a board certified radiologist [24], showing substantial differences in both the prevalence and sensitivity of the labels within each subclass (see Table 3). While this schema remains incomplete, even partial schema completion demonstrates substantial hidden stratification in this dataset.

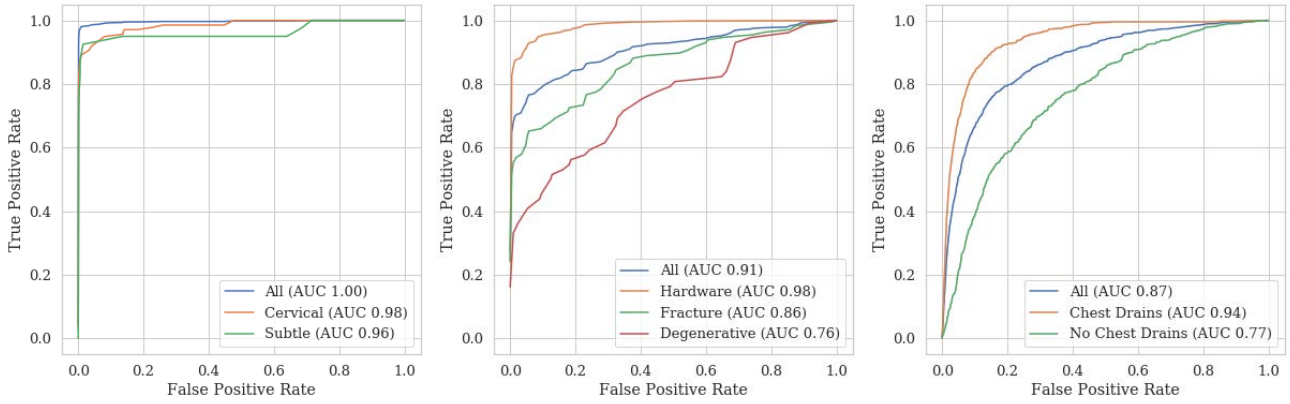
We hypothesize that the low label quality and subtle image features that characterize the degenerative joint disease subclass will result in reduced performance, and that the visually obvious metalwork subclass will have high performance (despite low prevalence). We train a DenseNet-169 on the normal/abnormal labels, with 13,942 cases used for training and 714 cases held-out for testing [26]. In Fig. 2(b), we present ROC curves and AUC values for each subclass and in aggregate. We observe that AUC for the easy-to-detect hardware subclass (0.98) is higher than aggregate AUC (0.91), despite the low subclass prevalence. As expected, we also find degraded AUC for degenerative disease (0.76), which has low-sensitivity superclass labels and subtle visual features (Table 3).

4.2 Error Auditing

We next use error auditing to show that the clinical utility of a common model for classifying the CXR-14 dataset is substantially reduced by existing hidden stratification effects in the pneumothorax class that result from the presence of a spurious correlate.

Subclass	Subclass Prevalence	Superclass Label Sensitivity
Fracture	0.30	0.92
Metalwork	0.11	0.85
DJD	0.43	0.60

Table 3: MURA "abnormal" label prevalence and sensitivity for the subclasses of "fracture," "metalwork," and "degenerative joint disease (DJD)." The degenerative joint disease subclass labels have the highest prevalence but the lowest sensitivity with respect to review by a board-certified radiologist.



(a) Adelaide Hip Abnormal **(b) MURA Abnormal** **(c) CXR14 Pneumothorax**

Figure 2: ROC curves for subclasses of the (a) abnormal Adelaide Hip Fracture superclass (b) abnormal MURA superclass and (c) pneumothorax CXR14 superclass. All subclass AUCs are significantly different than the overall task (DeLong p<0.05) for MURA and CXR14. For hip fracture, the AUCs themselves are not statistically different via a two-sided test (DeLong p>0.05), but the sensitivities are statistically different (p<0.01) at the relevant operating point [15]—see Table 2 for details. For MURA, sensitivities at 0.50 specificity are 0.93 (All), 1.00 (Hardware), 0.89 (Fracture), 0.80 (Degenerative). For CXR14, sensitivities at 0.50 specificity are 0.94 (All), 0.99 (Drain), and 0.85 (No Drain). For hip fracture, sensitivities at 0.50 specificity are 1.00 (All), 1.00 (Cervical), and 0.95 (Subtle)

CXR-14: The CXR-14 dataset is a large-scale dataset for pathology detection in chest radiographs [35]. This dataset was released in 2017 and updated later the same year, containing 112,120 frontal chest films from 30,805 unique patients. Each image was labeled for one of 14 different thoracic pathologies. In our analysis, we leverage a pretrained DenseNet-121 model provided by Zech [40] which reproduces the procedure and results of Rajpurkar et al. [27] on this dataset.

During error auditing, where examples of false positive and false negative predictions from the pretrained model were visually reviewed by a board certified radiologist [24], it was observed that pneumothorax cases without chest drains were highly prevalent (i.e., enriched) in the false negative class. A chest drain is a non-causal image feature in the setting of pneumothorax, as this device is the common form of treatment for the condition. As such, not only does this reflect a spurious correlate, but the correlation is in fact highly clinically relevant; untreated pneumothoraces are life-threatening while treated pneumothoraces are benign. To explore this audit-detected stratification, pneumothorax subclass labels for "chest drain" and "no chest drain" were provided by a board-certified radiologist (LOR) for each element of the test set. Due to higher prevalence of scans with chest drains in the dataset, clear discriminative features of a chest drain, and high label quality for the scans with chest drains, we hypothesize that a model trained on the CXR-14 dataset will attain higher performance on the pneumothorax subclass with chest drains than that without chest drains.

We present ROC curves for each pneumothorax subclass in Fig. 2(c). While overall pneumothorax ROC-AUC closely matches that reported in Rajpurkar et al. [28] at 0.87, pneumothorax ROC-AUC was 0.94 on the subclass with chest drains, but only 0.77 on the subclass without chest drains. We find that 80% of pneumothoraces in the test set contained a chest drain, and that positive predictive value on this subset was 30% higher (0.90) than on those with no chest drain (0.60). These results suggest that clearly identifiable spurious correlates can also cause clinically important hidden stratification.

4.3 Algorithmic Approaches: Unsupervised Clustering

While schema completion and error auditing have allowed us to identify hidden stratification problems in multiple medical machine learning datasets, each requires substantial effort from clinicians. Further, in auditing there is no guarantee that an auditor will recognize underlying patterns in the model error profile. In this context, unsupervised learning techniques can be valuable tools in automatically identifying hidden stratification. We show that even simple k-means clustering can detect several of the hidden subsets identified above via time-consuming human review or annotation.

For each superclass, we apply k-means clustering to the pre-softmax feature vector of all test set examples within that superclass using $k \in \{2, 3, 4, 5\}$. For each value of k, we select the two clusters with greater than 100 constituent points that have the largest difference in error rates (to select a "high error cluster" and "low error cluster" for each k). Finally, we return the pair of high and low error clusters that have the largest Euclidean distance between their centroids. Ideally, examining these high and low error clusters would help human analysts identify salient stratifications in the data. Note that our clustering hyperparameters were coarsely tuned, and could likely be improved in practice.

Dataset-Superclass (Subclass)	Difference in Subclass Prevalence (High Error Cluster, Low Error Cluster)	Overall Subclass Prevalence
CXR14-Pneumothorax (Drains)	0.68 (0.17, 0.84)	0.80
CIFAR-Carnivores (Bears)	0.30 (0.36, 0.06)	0.20
CIFAR-Outdoor (Forest)	0.28 (0.36, 0.08)	0.20
CIFAR-Household (Lamp)	0.16 (0.28, 0.12)	0.20
MURA-Abnormal (Hardware)	0.03 (0.29, 0.26)	0.11
MURA-Abnormal (Degenerative)	0.04 (0.12, 0.08)	0.43

Table 4: Subclass prevalence in high and low error clusters on CIFAR, MURA, and CXR14.

To demonstrate the potential utility of this approach, we apply it to several datasets analyzed above, and report results in Table 4. We find that while this simple k-means clustering approach does not always yield meaningful separation (e.g. on MURA), it does produce clusters with a high proportion of drains on CXR-14 and a high proportion of various high-error classes (bear, forest, lamp) on CIFAR-100. In practice, such an approach could be used both to assist human auditors in identifying salient stratifications in the data and to confirm that schema completion has been successful.

5 DISCUSSION

We find that hidden stratification can lead to markedly different superclass and subclass performance when labels for the subclasses have different levels of accuracy, when the subclasses are imbalanced, when discriminative visual features are subtle, or when spurious correlates such as chest drains are present. We observe these trends on both a controlled CIFAR-100 environment and multiple clinical datasets.

The clinical implications of hidden stratification will vary by task. Our MURA results, for instance, are unlikely to be clinically relevant, because degenerative disease is rarely a significant or unexpected finding, nor are rapid complications likely. We hypothesize that labels derived from clinical practice are likely to demonstrate this phenomenon; that irrelevant or unimportant findings are often elided by radiologists, leading to reduced label quality for less significant findings.

The findings in the CXR14 task are far more concerning. The majority of x-rays in the pneumothorax class contain chest drains, the presence of which is a healthcare process variable that is not causally linked to pneumothorax diagnosis. Importantly, the presence of a chest drain means these pneumothorax cases are already treated and are therefore at almost no risk of pneumothorax-related harm. In this experiment, we see that the performance in the clinically important subclass of cases without chest drains is far worse than the primary task results would suggest. We could easily imagine a situation where a model is justified for clinical use or regulatory approval with the results from the primary task alone, as the images used for testing simply reflect the clinical set of patients with pneumothoraces.

While this example is quite extreme, this does correspond with the medical truism that serious disease is typically less common than non-serious disease. These results suggest that image analysis systems that appear to perform well on a given task may fail to identify the most clinically important cases. This behavior is particularly concerning when comparing these systems to human experts, who focus a great deal of effort on specifically learning to identify rare, dangerous, and subtle disease variants.

The performance of medical image analysis systems is unlikely to be fully explained by the prevalence and accuracy of the labels, or even the dataset size. In the MURA experiment (see Figure 2), the detection of metalwork is vastly more accurate than the detection of fractures or degenerative change, despite this subclass being both smaller and less accurately labelled than fractures. We hypothesize that the nature of the visual features is important as well; metalwork is highly visible and discrete, as metal is significantly more dense (with higher pixel values) than any other material on x-ray. While our understanding of what types of visual features are more learnable than others is limited, it is not unreasonable to assume that detecting metal in an x-ray is far easier for a deep learning model than identifying a subtle fracture (and particularly on downsampled images). Similarly, chest drains are highly recognizable in pneumothorax imaging, and small untreated pneumothoraces are subtle enough to be commonly missed by radiologists. It is possible that this effect exaggerates the discrepancy in performance on the pneumothorax detection task, beyond the effect of subclass imbalance alone. Finally, it is worth noting that there will likely be stratifications within a dataset that are *not* distinguishable by imaging, meaning that the testing for hidden stratification is likely a necessary, but not sufficient condition for models that perform in a clinically optimal manner.

We show that a simple unsupervised approach to identify unrecognized subclasses often produces clusters containing different proportions of cases from the hidden subclasses our analysis had previously identified. While these results support other findings that demonstrate the utility of hidden-state clustering in model development [20], the relatively simple technique presented here should be considered only a first attempt at unsupervised identification of hidden stratification [5, 32]. Indeed, it remains to be seen if these automatically produced clusters can be useful in practice, either for finding clinically important subclasses or for use in retraining image analysis models for improved subclass performance, particularly given the failure of this method in the detection of clinically relevant subclasses in the MURA task. More advanced semi-supervised methods such as those of [8] may ultimately be required to tackle this problem, or it may be the case that both unsupervised and semi-supervised approaches are unable to contribute substantially, leaving us reliant on time-consuming methodical human review. Importantly, our experiments are limited in that they

do not explore the full range of medical image analysis tasks, so the results will have variable applicability to any given scenario. The findings presented here are intended specifically to highlight the largely underappreciated problem of hidden stratification in clinical imaging datasets, and to suggest that awareness of hidden stratification is important and should be considered (even if to be dismissed) when planning, building, evaluating, and regulating clinical image analysis systems.

6 CONCLUSION

Hidden stratification in medical image datasets appears to be a significant and underappreciated problem. Not only can the unrecognized presence of hidden subclasses lead to impaired subclass performance, but this may even result in unexpected negative clinical outcomes in situations where image analysis models silently fail to identify serious but rare, noisy, or visually subtle subclasses. Acknowledging the presence of visual variation within class labels is likely to be important when building and evaluating the next generation of medical image analysis systems. Indeed, our results suggest that models should not be certified for deployment by regulators unless careful testing for hidden stratification has been performed. While this will require substantial effort from the community, bodies such as professional organizations, academic institutions, and national standards boards can help ensure that we can leverage the enormous potential of machine learning in medical imaging without causing patients harm as a result of hidden stratification effects in our models.

REFERENCES

[1] Denis Agniel, Isaac S Kohane, and Griffin M Weber. 2018. Biases in electronic health record data due to processes within the healthcare system: retrospective observational study. *BMJ* 361 (April 2018), k1479.

[2] Marcus A Badgeley, John R Zech, Luke Oakden-Rayner, Benjamin S Glicksberg, Manway Liu, William Gale, Michael V McConnell, Bethany Percha, Thomas M Snyder, and Joel T Dudley. 2019. Deep learning predicts hip fracture using confounding patient and healthcare variables. *NPJ Digit Med* 2 (April 2019), 31.

[3] Nicholas Bien, Pranav Rajpurkar, Robyn L Ball, Jeremy Irvin, Allison Park, Erik Jones, Michael Bereket, Bhavik N Patel, Kristen W Yeom, Katie Shpanskaya, Safwan Halabi, Evan Zucker, Gary Fanton, Derek F Amanatullah, Christopher F Beaulieu, Geoffrey M Riley, Russell J Stewart, Francis G Blankenberg, David B Larson, Ricky H Jones, Curtis P Langlotz, Andrew Y Ng, and Matthew P Lungren. 2018. Deep-learning-assisted diagnosis for knee magnetic resonance imaging: Development and retrospective validation of MRNet. *PLoS Med.* 15, 11 (Nov. 2018), e1002699.

[4] Mateusz Buda, Atsuto Maki, and Maciej A Mazurowski. 2018. A systematic study of the class imbalance problem in convolutional neural networks. *Neural Netw.* 106 (Oct. 2018), 249–259.

[5] Tadeusz Caliński and Jerzy Harabasz. 1974. A dendrite method for cluster analysis. *Communications in Statistics-Theory and Methods* 3, 1 (1974), 1–27.

[6] Gabriele Campanella, Matthew G Hanna, Luke Geneslaw, Allen Miraflor, Vitor Werneck Krauss Silva, Klaus J Busam, Edi Brogi, Victor E Reuter, David S Klimstra, and Thomas J Fuchs. 2019. Clinical-grade computational pathology using weakly supervised deep learning on whole slide images. *Nature Medicine* 25, 8 (2019), 1301–1309.

[7] Lon R Cardon and Lyle J Palmer. 2003. Population stratification and spurious allelic association. *Lancet* 361, 9357 (2003), 598–604.

[8] Vincent Chen, Sen Wu, Alexander J Ratner, Jen Weng, and Christopher Ré. 2019. Slice-based learning: A programming model for residual learning in critical data slices. In *Advances in neural information processing systems*. 9392–9402.

[9] Sasank Chilamkurthy, Rohit Ghosh, Swetha Tanamala, Mustafa Biviji, Norbert G Campeau, Vasantha Kumar Venugopal, Vidur Mahajan, Pooja Rao, and Prashant Warier. 2018. Deep learning algorithms for detection of critical findings in head CT scans: a retrospective study. *Lancet* 392, 10162 (Dec. 2018), 2388–2396.

[10] Jared Dunnmon, Alexander Ratner, Nishith Khandwala, Khaled Saab, Matthew Markert, Hersh Sagreiya, Roger Goldman, Christopher Lee-Messer, Matthew Lungren, Daniel Rubin, and Christopher Ré. 2019. Cross-Modal Data Programming Enables Rapid Medical Machine Learning. *arXiv preprint arXiv: 1903.11101* (March 2019).

[11] Jared A Dunnmon, Darvin Yi, Curtis P Langlotz, Christopher Ré, Daniel L Rubin, and Matthew P Lungren. 2019. Assessment of Convolutional Neural Networks for Automated Classification of Chest Radiographs. *Radiology* 290, 2 (Feb. 2019), 537–544.

[12] Andre Esteva, Brett Kuprel, Roberto A Novoa, Justin Ko, Susan M Swetter, Helen M Blau, and Sebastian Thrun. 2017. Dermatologist-level classification of skin cancer with deep neural networks. *Nature* 542, 7639 (Feb. 2017), 115–118.

[13] Andre Esteva, Alexandre Robicquet, Bharath Ramsundar, Volodymyr Kuleshov, Mark DePristo, Katherine Chou, Claire Cui, Greg Corrado, Sebastian Thrun, and Jeff Dean. 2019. A guide to deep learning in healthcare. *Nature Medicine* 25, 1 (2019), 24.

[14] Jason A Fries, Paroma Varma, Vincent S Chen, Ke Xiao, Heliodoro Tejeda, Priyanka Saha, Jared Dunnmon, Henry Chubb, Shiraz Maskatia, Madalina Fiterau, Scott Delp, Euan Ashley, Christopher Ré, and James R Priest. 2019. Weakly supervised classification of aortic valve malformations using unlabeled cardiac MRI sequences. *Nat. Commun.* 10, 1 (July 2019), 3111.

[15] William Gale, Luke Oakden-Rayner, Gustavo Carneiro, Andrew P Bradley, and Lyle J Palmer. 2017. Detecting hip fractures with radiologist-level performance using deep neural networks. *arXiv preprint arXiv:1711.06504* (2017).

[16] Varun Gulshan, Lily Peng, Marc Coram, Martin C Stumpe, Derek Wu, Arunachalam Narayanaswamy, Subhashini Venugopalan, Kasumi Widner, Tom Madams, Jorge Cuadros, Ramasamy Kim, Rajiv Raman, Philip C Nelson, Jessica L Mega, and Dale R Webster. 2016. Development and Validation of a Deep Learning Algorithm for Detection of Diabetic Retinopathy in Retinal Fundus Photographs. *JAMA* 316, 22 (Dec. 2016), 2402–2410.

[17] Holger A Haenssle, Christine Fink, R Schneiderbauer, Ferdinand Toberer, Timo Buhl, A Blum, A Kalloo, A Ben Hadj Hassen, L Thomas, A Enk, and Others. 2018. Man against machine: diagnostic performance of a deep learning convolutional neural network for dermoscopic melanoma recognition in comparison to 58 dermatologists. *Ann. Oncol.* 29, 8 (2018), 1836–1842.

[18] Jeremy Irvin, Pranav Rajpurkar, Michael Ko, Yifan Yu, Silviana Ciurea-Ilcus, Chris Chute, Henrik Marklund, Behzad Haghgoo, Robyn Ball, Katie Shpanskaya, and Others. 2019. Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison. *arXiv preprint arXiv:1901. 07031* (2019).

[19] Alex Krizhevsky, Vinod Nair, and Geoffrey Hinton. 2009. Cifar-10 and cifar-100 datasets. *URl: https://www. cs. toronto. edu/kriz/cifar. html* 6 (2009).

[20] Jiamin Liu, Jianhua Yao, Mohammadhadi Bagheri, Veit Sandfort, and Ronald M Summers. 2019. A Semi-Supervised CNN Learning Method with Pseudo-class Labels for Atherosclerotic Vascular Calcification Detection. *2019 IEEE 16th International Symposium on Biomedical Imaging (ISBI 2019)* (2019).

[21] Vidur Mahajan, Vasanthakumar Venugopal, Saumya Gaur, Salil Gupta, Murali Murugavel, and Harsh Mahajan. 2019. The Algorithmic Audit: Working with Vendors to Validate Radiology-AI Algorithms - How We Do It. *viXra* (July 2019).

[22] Maciej A Mazurowski, Piotr A Habas, Jacek M Zurada, Joseph Y Lo, Jay A Baker, and Georgia D Tourassi. 2008. Training neural network classifiers for medical decision making: the effects of imbalanced datasets on classification performance. *Neural Netw.* 21, 2-3 (March 2008), 427–436.

[23] Stephanie A Mulherin and William C Miller. 2002. Spectrum bias or spectrum effect? Subgroup variation in diagnostic test evaluation. *Annals of Internal Medicine* 137, 7 (2002), 598–602.

[24] Luke Oakden-Rayner. 2020. Exploring Large-scale Public Medical Image Datasets. *Academic Radiology* 27, 1 (2020), 106–112.

[25] Inioluwa Deborah Raji, Andrew Smart, Rebecca N White, Margaret Mitchell, Timnit Gebru, Ben Hutchinson, Jamila Smith-Loud, Daniel Theron, and Parker Barnes. 2020. Closing the AI Accountability Gap: Defining an End-to-End Framework for Internal Algorithmic Auditing. *arXiv preprint arXiv:2001.00973* (2020).

[26] Pranav Rajpurkar, Jeremy Irvin, Aarti Bagul, Daisy Ding, Tony Duan, Hershel Mehta, Brandon Yang, Kaylie Zhu, Dillon Laird, Robyn L Ball, et al. 2017. Mura: Large dataset for abnormality detection in musculoskeletal radiographs. *arXiv preprint arXiv:1712.06957* (2017).

[27] Pranav Rajpurkar, Jeremy Irvin, Robyn L Ball, Kaylie Zhu, Brandon Yang, Hershel Mehta, Tony Duan, Daisy Ding, Aarti Bagul, Curtis P Langlotz, Bhavik N Patel, Kristen W Yeom, Katie Shpanskaya, Francis G Blankenberg, Jayne Seekins, Timothy J Amrhein, David A Mong, Safwan S Halabi, Evan J Zucker, Andrew Y Ng, and Matthew P Lungren. 2018. Deep learning for chest radiograph diagnosis: A retrospective comparison of the CheXNeXt algorithm to practicing radiologists. *PLoS Med.* 15, 11 (Nov. 2018), e1002686.

[28] Pranav Rajpurkar, Jeremy Irvin, Kaylie Zhu, Brandon Yang, Hershel Mehta, Tony Duan, Daisy Ding, Aarti Bagul, Curtis Langlotz, Katie Shpanskaya, Matthew Lungren, and Andrew Ng. 2017. CheXNet: Radiologist-level pneumonia detection on chest x-rays with deep learning. *arXiv preprint arXiv:1711.05225* (2017).

[29] Alexander J Ratner, Henry Ehrenberg, Zeshan Hussain, Jared Dunnmon, and Christopher Ré. 2017. Learning to Compose Domain-Specific Transformations for Data Augmentation. In *Advances in Neural Information Processing Systems 30*, I Guyon, U V Luxburg, S Bengio, H Wallach, R Fergus, S Vishwanathan, and R Garnett (Eds.). Curran Associates, Inc., 3236–3246.

[30] Benjamin Recht, Rebecca Roelofs, Ludwig Schmidt, and Vaishaal Shankar. 2018. Do CIFAR-10 classifiers generalize to CIFAR-10? *arXiv preprint arXiv:1806.00451* (2018).

[31] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. Why should i trust you?: Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, 1135–1144.

[32] Peter J Rousseeuw. 1987. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *J. Comput. Appl. Math.* 20 (1987), 53–65.

[33] Andrew D Selbst. 2017. Disparate impact in big data policing. *Ga. L. Rev.* 52 (2017), 109.

[34] Pu Wang, Tyler M Berzin, Jeremy Romek Glissen Brown, Shishira Bharadwaj, Aymeric Becq, Xun Xiao, Peixi Liu, Liangping Li, Yan Song, Di Zhang, Yi Li, Guangre Xu, Mengtian Tu, and Xiaogang Liu. 2019. Real-time automatic detection system increases colonoscopic polyp and adenoma detection rates: a prospective randomised controlled study. *Gut* (Feb. 2019).

[35] Xiaosong Wang, Yifan Peng, Le Lu, Zhiyong Lu, Mohammadhadi Bagheri, and Ronald M Summers. 2017. Chestx-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. In *IEEE Conference on CVPR, Computer Vision and Pattern Recognition (2017)*. 3462–3471.

[36] Julia K Winkler, Christine Fink, Ferdinand Toberer, Alexander Enk, Teresa Deinlein, Rainer Hofmann-Wellenhof, Luc Thomas, Aimilios Lallas, Andreas Blum, Wilhelm Stolz, and Holger A Haenssle. 2019. Association Between Surgical Skin Markings in Dermoscopic Images and Diagnostic Performance of a Deep Learning Convolutional Neural Network for Melanoma Recognition. *JAMA Dermatology* (2019).

[37] Saining Xie, Ross Girshick, Piotr Dollár, Zhuowen Tu, and Kaiming He. 2016. Aggregated Residual Transformations for Deep Neural Networks. *arXiv preprint arXiv:1611.05431* (Nov. 2016).

[38] Wei Yang. 2019. pytorch-classification. https://github.com/bearpaw/pytorch-classification

[39] Bangpeng Yao, Aditya Khosla, and Li Fei-Fei. 2011. Combining randomization and discrimination for fine-grained image categorization. In *CVPR 2011*. IEEE, 1577–1584.

[40] John Zech. 2019. reproduce-chexnet. https://github.com/jrzech/reproduce-chexnet

[41] John R Zech, Marcus A Badgeley, Manway Liu, Anthony B Costa, Joseph J Titano, and Eric K Oermann. 2018. Confounding variables can degrade generalization performance of radiological deep learning models. *arXiv preprint arXiv:1807.00431* (July 2018).

Section 2: Issues with testing

The preclinical testing of medical imaging AI systems typically involves some combination of three elements [1,2,30]:

1. a measurement of model performance on "hold-out" data; images and patients which have not been seen by the algorithm during model training
2. a reader study, comparing human performance against AI performance on all or a subset of this hold out data
3. an external validation testing the AI model on independent data from a different geographic location

The ideas of 'hold-out' data and external validations are commonly discussed in the medical machine learning literature, and can be interpreted as attempts to reduce the implementation gap. Image analysis AI models can learn unintended cues in the training data [31]; image features that correlate with the clinical target in the training dataset but are not causally related and will not be useful outside of that environment. For example, AI models have been shown to learn that radiographic markers are associated with pneumonia or that surgical skin markings are a sign of melanoma [11,13], but as these features are related to local clinical processes rather than the diseases themselves, use of these models can result in unexpectedly poor performance in settings where this correlation is not preserved.

By testing the model on unseen, independent data, particularly if those data come from a different geographic site (with a different patient population and different clinical processes), it is reasonable to assume that many of these unintended cues will no longer be useful for the task. Thus, we can obtain a more realistic estimate of real-world model AI performance and reduce the implementation gap.

Another common element of medical AI studies is a comparison against human experts. A reader study can provide a good clinical baseline for any task which humans have expertise in, however it is necessary to recognise that achieving equivalent or better performance than humans does not *necessarily* imply that there is no implementation gap. The capability of AI models to learn unintended cues means that models perform the task *differently* than a human, who would for example never mistake a radiographic marker for a feature of pneumonia. While unintended cue learning often can be detected as poor performance during testing, it can also result in *equal but different* performance, where the same number of errors are made but the specific errors are different, affecting different patients and potentially producing different outcomes. Given the fact that some medical errors are higher risk than others, it is entirely possible for AI systems to produce fewer errors than human experts while also increasing the overall risk to patients.

2.1 Metrics and human performance baselines

The reader study is the most common study design for medical AI testing prior to regulatory approval. This is referenced in regulatory guidelines [32], study design checklists [33], and can be clearly seen in a review of the literature; almost all medical AI studies in a number of systematic reviews have relied upon a multi-reader baseline as a comparison with current clinical practice [1,2].

The analysis of reader studies using receiver operating characteristic (ROC) curves is widespread in diagnostic radiology and the assessment of diagnostic test accuracy [34–36]. These methods have tended to be applied in a particular set of clinical scenarios however, particularly breast imaging where cases are scored for the probability of malignancy, and the extension of the underlying assumptions to typical medical AI evaluation use cases (where these scores are unavailable) has been problematic. When disease scores are unavailable and the clinical task is treated as a binary prediction, as is the case in most diagnostic tasks in medical imaging, it is impossible to construct reader-wise ROC curves due to the absence of multiple operating points per reader.

In this setting, the standard practice has consisted of reporting human performance with the average of sensitivity and specificity, independently pooled across readers, which systematically under-estimates human performance and in doing so can overestimate the clinical value of AI systems.

In "**Docs are ROCs: A simple off-the-shelf approach for estimating average human performance in diagnostic studies"** [39] I present a methodologically justified approach to the analysis of reader data in the most common AI study designs, where we compare human and AI performance on one or a set of binary tasks. By utilising well validated methods from the field of statistical meta-analysis for diagnostic test accuracy studies, we can better estimate human performance and produce fairer clinical baselines against which to compare AI models.

Docs are ROCs: A simple off-the-shelf approach for estimating average human performance in diagnostic studies

Publication status: Unpublished, preprinted at *arXiv* (https://arxiv.org/abs/2009.11060), 2020.

Contribution: 80%

Detailed description contribution: I was responsible for planning, study design, experimentation, and drafting/editing the manuscript.

Senior author Professor Lyle Palmer was involved in planning and study design as well as manuscript editing, and contributed an estimated 20% of the work.

Certification from co-authors:

Professor Lyle Palmer

Declaration: This publication was part of the work undertaken during and for my HDR candidature.

Docs are ROCs: a simple off-the-shelf approach for estimating average human performance in diagnostic studies

Dr Luke Oakden-Rayner, Dr Lyle Palmer

The Australian Institute for Machine Learning, The University of Adelaide, Australia.
luke.oakden-rayner@adelaide.edu.au
lyle.palmer@adelaide.edu.au

Introduction

Sensitivity and specificity are among the most common and important metrics in diagnostic medical research, favoured for their ability to summarise both false positive and false negative errors[1]. Their invariance to disease prevalence allows for a direct numerical comparison of performance across tests or sites with different rates of disease. Unsurprisingly, these metrics are widely reported in medical artificial intelligence (AI) studies, particularly where human performance is compared to that of AI models in multi-reader multi-case (MRMC) study designs. Given the variability in performance between readers on any given set of cases, multiple human readers are required to estimate the range of "average" human performance, but there is no well-motivated consensus on how to perform this averaging operation. A recent systematic review[2] noted that 'naive' averages of human sensitivity and/or specificity, or other metrics derived from these values, were reported in at least 70% of publications that compared human performance to AI models, a practice which is highly problematic.

The use of sensitivity and specificity to describe the discriminative performance of individual tests or readers is appropriate, but averaging these highly correlated metrics independently of each other is strongly discouraged in other, more methodologically mature domains such as the meta-analysis of diagnostic test accuracy studies. For example in "Guidelines for Meta-Analyses Evaluating Diagnostic Tests"[3] the authors write "In general, estimating mean sensitivity and specificity separately underestimates test accuracy". Gatsonis and Paliwal[4] even go as far as to say "the use of simple or weighted averages of sensitivity and specificity to draw statistical conclusions is not methodologically defensible." Similarly, the Cochrane handbook recommends these metrics be addressed together rather than in isolation when summarising the accuracy of a diagnostic test[5].

Despite these recommendations from reputable authors and bodies, the independent pooling of sensitivity and specificity (or use of similar pooled metrics such as the average F1 score and average accuracy) remain popular in the medical AI literature. Unfortunately, not only do these methods consistently underestimate human diagnostic performance, but they can bias the

experimental conclusions because the performance of the AI models are not similarly underestimated.

An alternative to the various pooled metrics is to treat the estimation of human performance as a bivariate modelling problem, operating on the justified assumption that sensitivity and specificity are correlated across readers. This approach has become the mainstay of the meta-analysis of diagnostic test accuracy across the last 50 years, supported by an extensive body of literature on the development and validation of meta-analytic models. Indeed, meta-analysis is considered the highest level of experimental evidence in clinical medicine [6], in part because of the robustness of these techniques.

Meta-analysis for diagnostic test accuracy studies involves the production of summary receiver operating characteristic (SROC) curves. The specific methods to do this are well covered in other publications[7–10], but it is worth noting that these techniques are well understood and validated in the biostatistics community, and that software implementations of these methods are widely available in most common programming languages. Typically all that is needed is the 2x2 contingency table for each reader and the software can do the rest.

Fixed effects vs random effects models

Briefly, there are two main families of models used for meta-analysis and SROC curve development; the fixed-effects and random-effects models. In simple terms the fixed effect models assume that the only difference between tests (in this case, the readers) is due to a single source of variation; that it is the choice of readers alone that contributes to these differences. In random-effects models, sometimes called hierarchical or two-level models, an estimation of further test heterogeneity is included. In the setting of reader studies common further sources of test heterogeneity include intra-user variability and the assessment of different cases by each reader.

In general, random-effects models are recommended for the meta-analysis of diagnostic test accuracy studies[5]. Given the multiple sources of heterogeneity in MRMC studies, this recommendation appears appropriate in this context as well.

Related literature

In the case where the outcome is binary, a point in ROC space will be produced representing the sensitivity and specificity derived from a simple 2x2 table. However, in order to derive ROC curves at the level of individual readers, the MRMC literature has been strongly focused on the use of ordinal scoring systems such as those used in mammography[11]. In the scenario where a diagnostic score contains at least 5 levels it is reasonable to produce ROC curves for individual readers and then average these curves themselves, summarising the performance across readers[12]. These methods cannot be used when the clinical diagnosis is made with a binary

response (e.g., “yes there is cancer” vs “no there is not”), and attempts to extend these methods to binary data[13] have not seen widespread uptake. In fact, it has become accepted practice to shoehorn ordinal scoring systems into tasks normally reported with binary responses (e.g., applying a 100 point scoring system to lung nodule detection[14], despite the fact that clinical radiologists only ever report that nodules are present or absent). This approach has even been tacitly endorsed by regulatory bodies[15], but is critically limited by its failure to reproduce clinical practice, raising significant concerns about the clinical relevance of this testing and the possibility of misleading laboratory effects[16].

In the medical AI literature, Rajpurkar et al used a constrained spline approach to summarise performance by estimating AUC[17]. This method assumes a symmetrical ROC curve (which is an uncommon distribution of readers in clinical practice), and produces confidence intervals with a bootstrap across cases alone, therefore underestimating the standard errors of the AUC by failing to incorporate the variability across readers.

Why do SROC analysis of multi-reader studies?

Aside from the already stated improvements in accuracy and methodological defensibility, SROC analysis has a number of attractive features compared to other commonly used methods.

First, it allows for the estimation of a single metric (the area under the SROC curve, also known as the SROC AUC) which summarises the discriminative ability of readers. Comparison between reader groups or readers and AI models is significantly simplified compared to separate consideration of sensitivity and specificity or similar metrics.

Second, it allows us to produce valid confidence intervals. When sensitivity and specificity are pooled separately, the confidence intervals are almost always calculated using the number of cases but *not* the number of readers. SROC analysis automatically takes both elements of variation into account. Importantly in common experimental scenarios (where $n_{Observers}$ <10) the number of readers contributes strongly to the estimation of variation.

Third, it avoids the need to select an arbitrary or unnatural (i.e., one that will never occur in clinical practice) operating point. If we consider that the position of a human reader along an SROC curve is related to their “aggressiveness” or risk-aversion, then these quantities are not fixed, either between readers or for individuals. SROC analysis allows for more control of the selection of an operating point if this is needed, and allows comparisons without operating point selection if this is more appropriate.

Fourth, SROC analysis allows for visual presentation of results in a way that is easy to interpret. Side-by-side ROC curves are understandable at a glance while conveying a great deal of information about the discriminative performance of each decision maker, and the ability to plot confidence intervals allows for a useful visual summary of an experiment.

Fifth and finally, SROC analysis can allow for easy comparison of subsets of readers. Many studies have included both expert and non-expert readers, and presentation of these results can be difficult. Single summary points (pooled sensitivity and specificity) are unjustified, but colour-coding of all the readers can be visually overwhelming if $n_{Observers}$ is high. Producing SROC curves for each subset can allow for easy comparisons between groups, and comparisons of SROC AUC values are well motivated (particularly given the differing number of readers and different variance between these subgroups).

Methods

We present examples of this meta-analytic approach applied to a variety of heavily cited reports in the medical AI literature, re-evaluating the presented ROC curves and primary comparisons. For the majority of these papers, the data from these studies have been reproduced from the published figures (i.e., sensitivity and specificity were “eyeballed” for each reader), although Tschandl et al. provided the raw reader data for their experiments[18].

All statistical analysis was performed in R v3.6.2[19] . SROC analysis was undertaken with the mada package v0.5.8[20], using the proportional hazards model described by Holling et al[9]**.**

Results

Dermatologist-level classification of skin cancer with deep neural networks

Esteva et al[21] described a deep learning model trained to distinguish melanoma from non-melanomatous skin lesions, comparing the performance of the model against 22 dermatologists asked to decide if a skin lesion requires biopsy.

Esteva et al reported the average performance of the dermatologists by pooling sensitivity and specificity independently. This “average dermatologist” point was inside the ROC curve for the AI model. This figure was accompanied by the statement that the “CNN outperforms any dermatologist whose sensitivity and specificity point falls below the blue curve of the CNN”, although no specific statement was made about the “average” dermatologist.

In figure 1 we apply a random-effects model meta-analysis of the performance of the dermatologists, showing the benefits of treating sensitivity and specificity as correlated values. The average point is inside the summary ROC curve, and in fact is at the limit of the 95% confidence interval. The SROC curve appears to better reflect the desired goal of describing an

average dermatologist. With the curved distribution of the model as a guide, only 4 out of 22 dermatologists are “worse” than the average sens/spec point.

This approach not only produces a more justified summary of human performance, but the area under the SROC curve is directly comparable to the AUC of the AI model. The reported AUC of the AI model (0.94, CI not provided) is compared to the dermatologists (SROC AUC = 0.97, 95% CI 0.96 - 0.98).

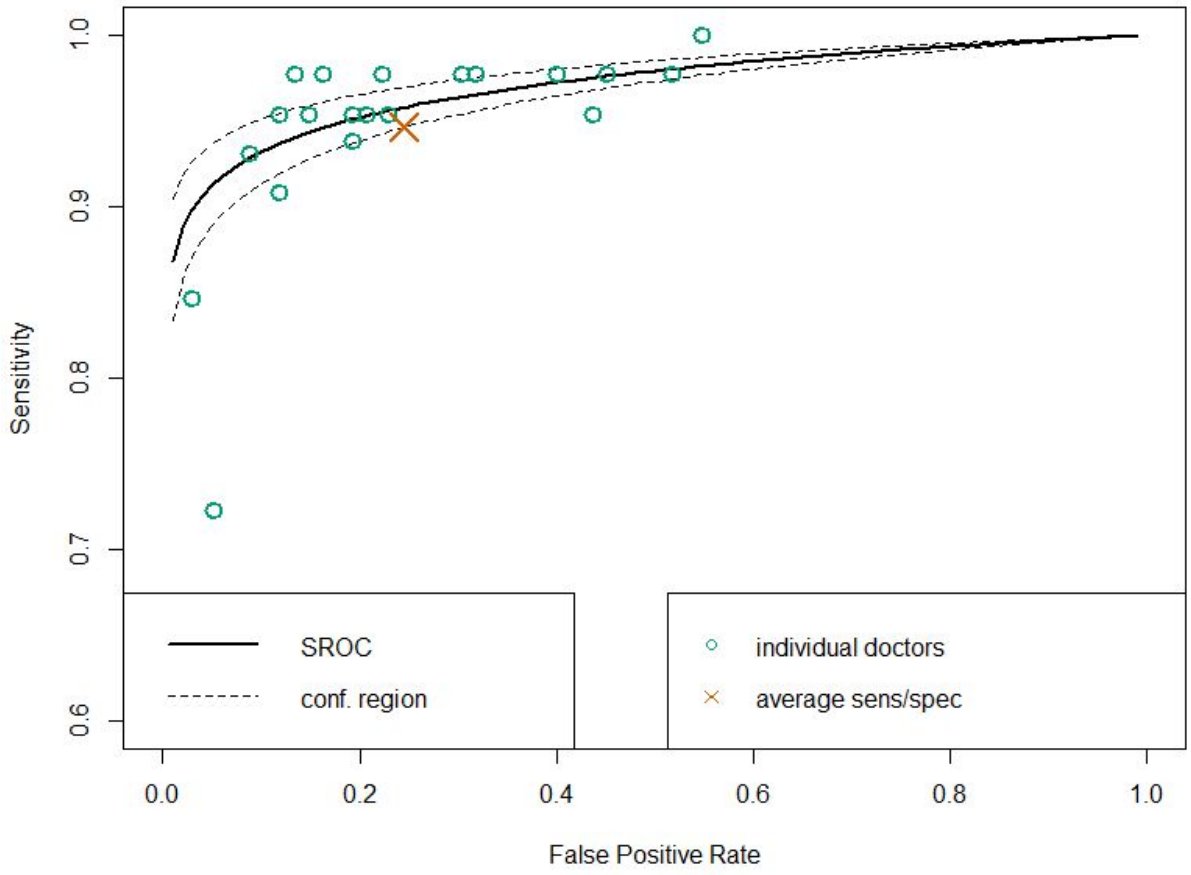


Figure 1: SROC analysis of Esteva et al[21] using a random effects model, demonstrating the individual performance of doctors (green circles), the average sensitivity and specificity of doctors (orange cross) and the SROC curve (black line) with associated 95% confidence region (dotted lines).

Deep-learning-assisted diagnosis for knee magnetic resonance imaging: Development and retrospective validation of MRNet

Bien and Rajpurkar et al[22] reported the comparison of a deep learning model against radiologists and orthopedic surgeons at the detection of meniscal tears, ACL tears, and combined for any abnormality. They reported the average performance of the dermatologists by pooling sensitivity and specificity independently.

In figure 2 we apply a random-effects meta-analysis to the performance of the clinical readers at the “any abnormality” task. Once again the “average” reader is below the SROC curve, and the SROC curve appears to be a more fair reflection of average reader performance.

The authors report that there was no significant difference between doctors and the AI model performance, albeit they allow for the fact that both the number of readers and number of cases are quite low leading to wide confidence intervals. In our approach, the AI model AUC of 0.937 (95% CI 0.895, 0.980) can be directly compared to the SROC AUC of 0.953 (95% CI 0.937, 0.969), which supports the statement from the authors.

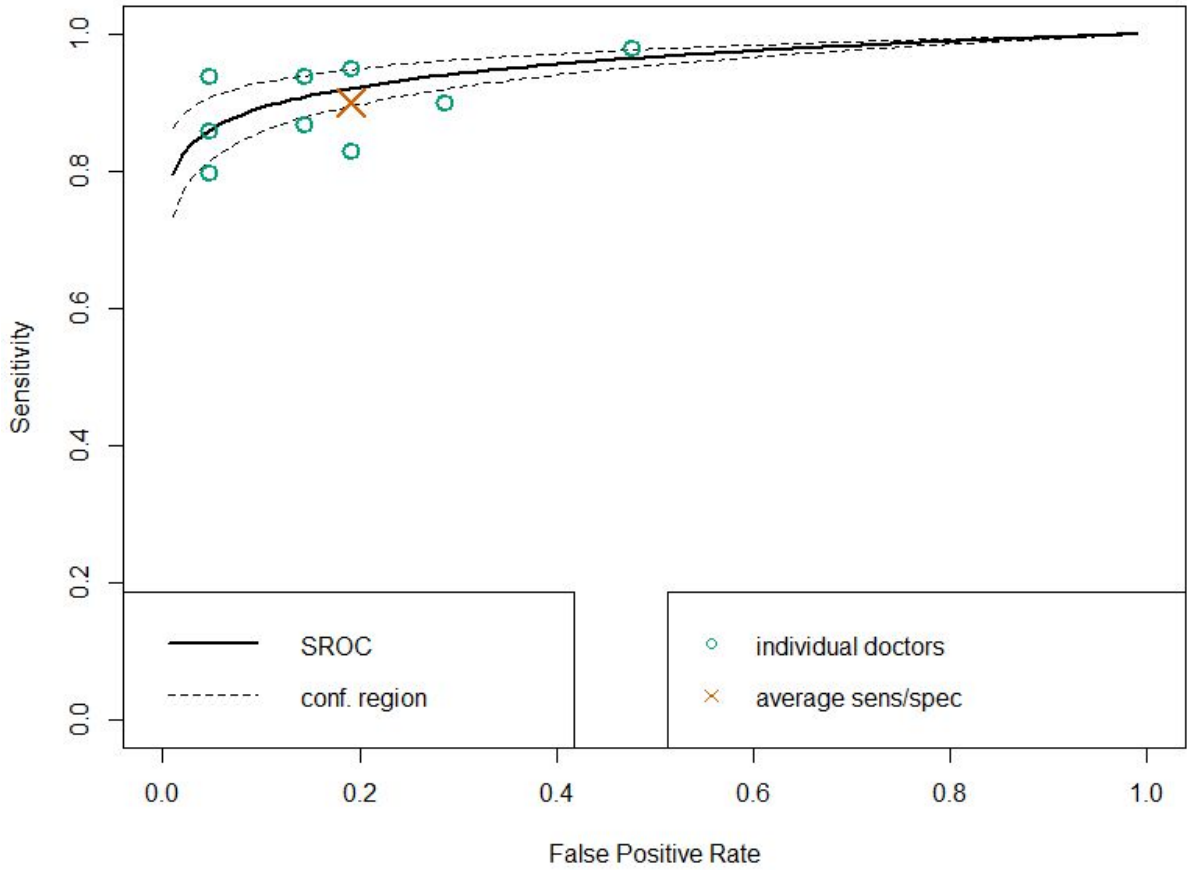


Figure 2: SROC analysis of Bien et al[22] using a random effects model, demonstrating the individual performance of doctors (green circles), the average sensitivity and specificity of

doctors (orange cross) and the SROC curve (black line) with associated 95% confidence region (dotted lines).

CheXNet: Radiologist-Level Pneumonia Detection on Chest X-Rays with Deep Learning

Rajpurkar and Irvin et al[23] compared the performance of an AI model against 4 radiologists at the task of pneumonia detection on chest x-ray. They initially reported the average of sensitivity and specificity for the radiologists, although the primary metric was changed to the average F1 score in a later revision.

In figure 3 we demonstrate a random-effects meta-analysis of human performance. In this case we see that the average sens/spec point is quite close to the SROC curve, but the example highlights another key benefit of this approach: the confidence intervals are very wide, due to the combination of a small test dataset (with only ~60 cases of pneumonia) and the small set of readers (n = 4). By failing to account for the latter source of variation, standard statistical tests based on the average sensitivity and specificity will be biased towards the alternative hypothesis. Rajpurkar et al report that the F1 score of the model is *significantly* better than the average F1 score of the radiologist, but the meta-analytic approach suggests that this is unlikely. While we cannot perform a null hypothesis test with the information provided in Rajpurkar et al, it can be appreciated that the evidence for a meaningful difference between the model AUC (0.77, CI not provided) and the radiologist SROC AUC (0.73, 95% CI 0.66, 0.83) is not compelling.

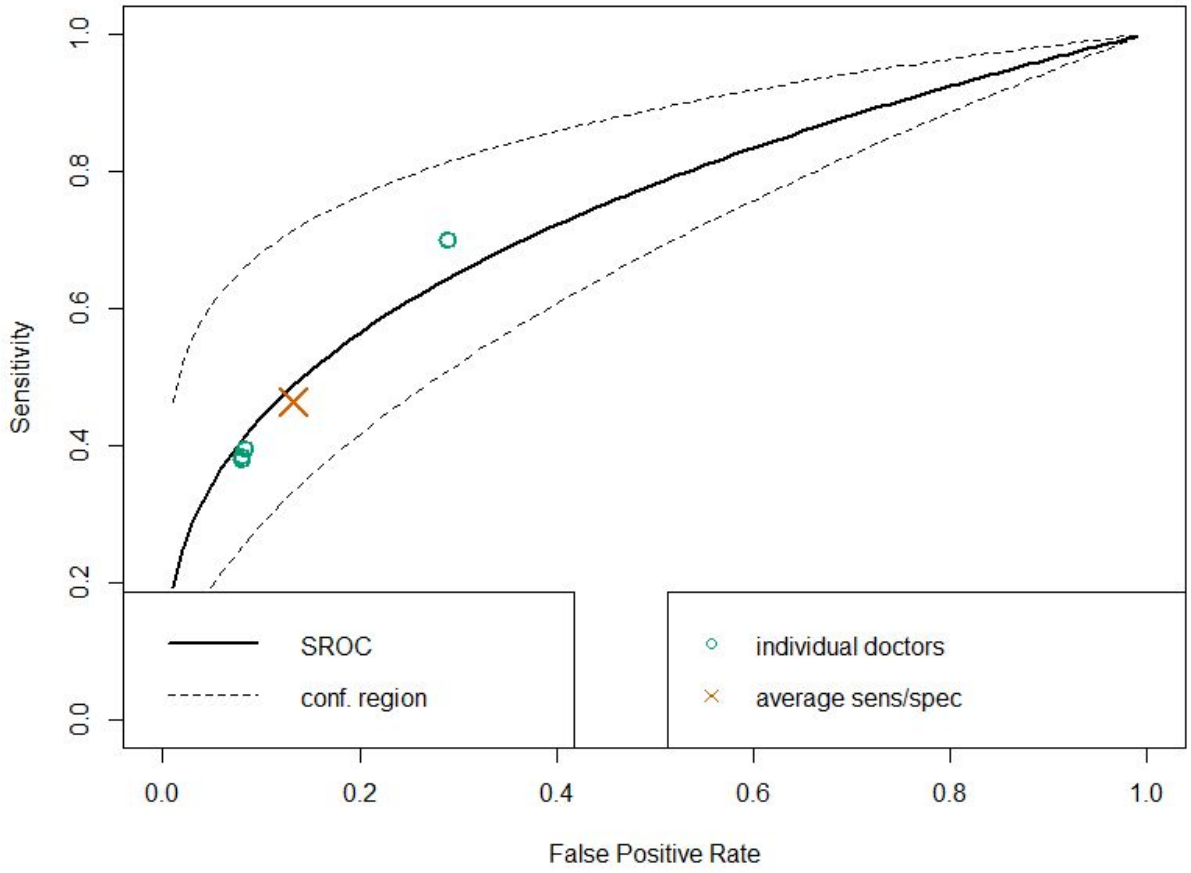


Figure 3: SROC analysis of Rajpurkar and Irvin et al[23] using a random effects model, demonstrating the individual performance of doctors (green circles), the average sensitivity and specificity of doctors (orange cross) and the SROC curve (black line) with associated 95% confidence region (dotted lines).

International evaluation of an AI system for breast cancer screening

McKinney et al[24] compared an AI model against radiologists for the detection of breast cancer at screening mammography. While mammography lends itself well to ordinal scoring, the authors also present results for a retrospective real-world dataset based on the binary decision of the readers with respect to the choice to perform a biopsy. Each reader read a different set of mammograms, each of different size.

In Figure 4 we demonstrate the use of a random-effects meta-analysis of human performance. The distribution of human readers is highly unusual, likely an artefact of the clinical demands of

mammography (where the false positive rates of readers are monitored to standardise biopsy rates).

The SROC curve again appears to capture a reasonable “average” performance more effectively than the average of sensitivity and specificity. In this example, the average point is below the 95% CI for the SROC curve, and is biased towards the bottom right of the set of readers. Only a small number of readers, who collectively reviewed an even smaller proportion of the overall cases, are inside the average point of sensitivity and specificity.

This example demonstrates the flexibility of SROC analysis. Not only does this method appropriately manage the unusual distribution of readers, but random-effect models can estimate the variability of cases between readers, accounting for sampling bias in the setting where each reader reviews different cases.

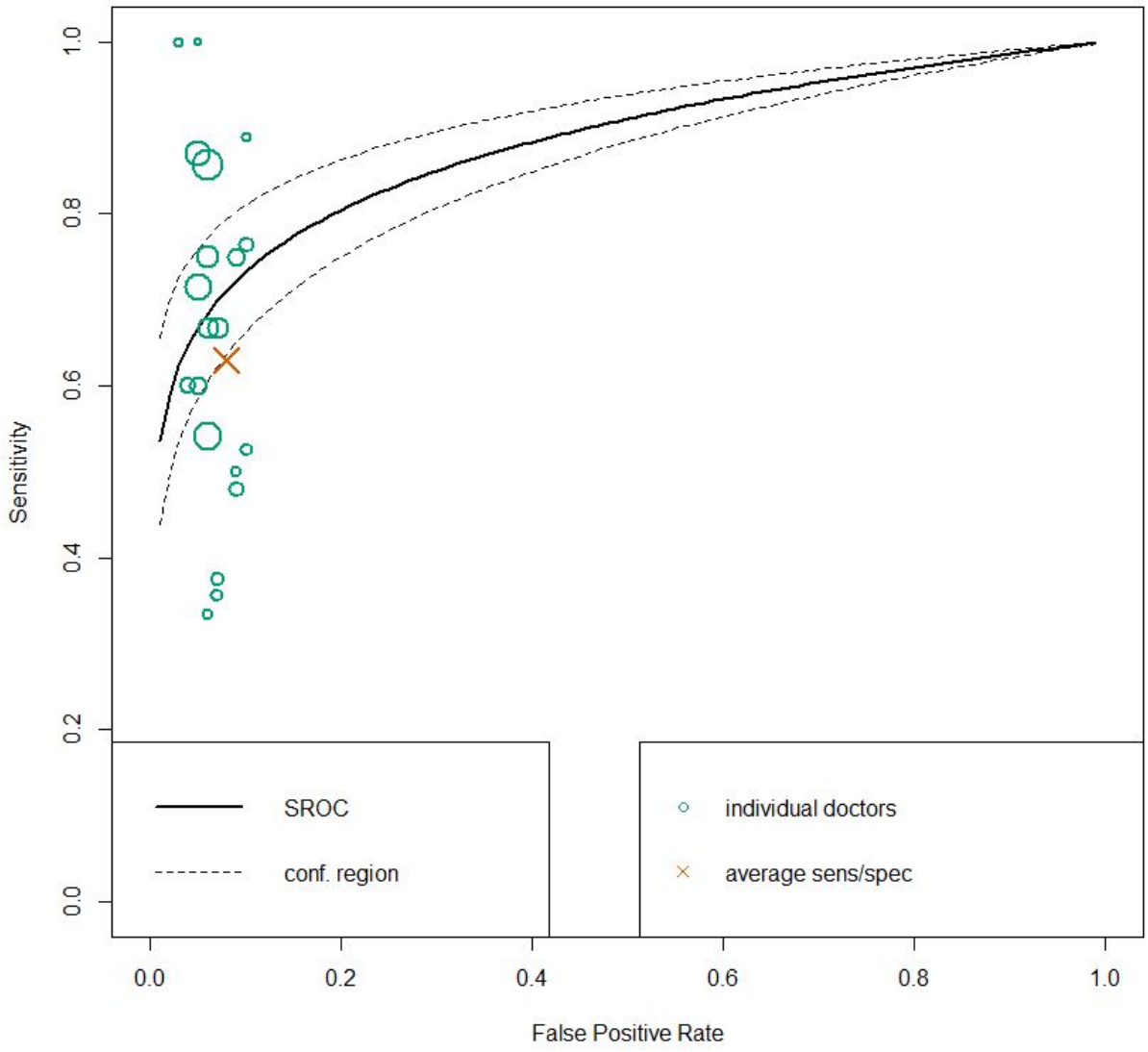


Figure 4: SROC analysis of McKinney et al[24] using a random effects model, demonstrating the individual performance of doctors (green circles), the average sensitivity and specificity of doctors (orange cross) and the SROC curve (black line) with associated 95% confidence region (dotted lines). The size of the green circles represents the number of cases each reader evaluated.

Human–computer collaboration for skin cancer recognition

Tschandl et al[18] report results for a 301 dermatologist reader study to classify lesions into benign and malignant categories, with each reader assessing 28 images. They report pooled average sensitivity, specificity, and several other similar statistics including the positive and

negative predictive values, and the youden J statistic. Notably, there was a wide range of experience levels among the readers, ranging from less than 1 year (n = 48) up to greater than 10 years (n = 15).

The extremely large number of readers are difficult to visualise on a single plot (figure 5a), however SROC analysis can greatly improve the visibility of subgroup comparisons (figure 5b). Again, we notice that the “average” sensitivity and specificity points are well below the respective curves.

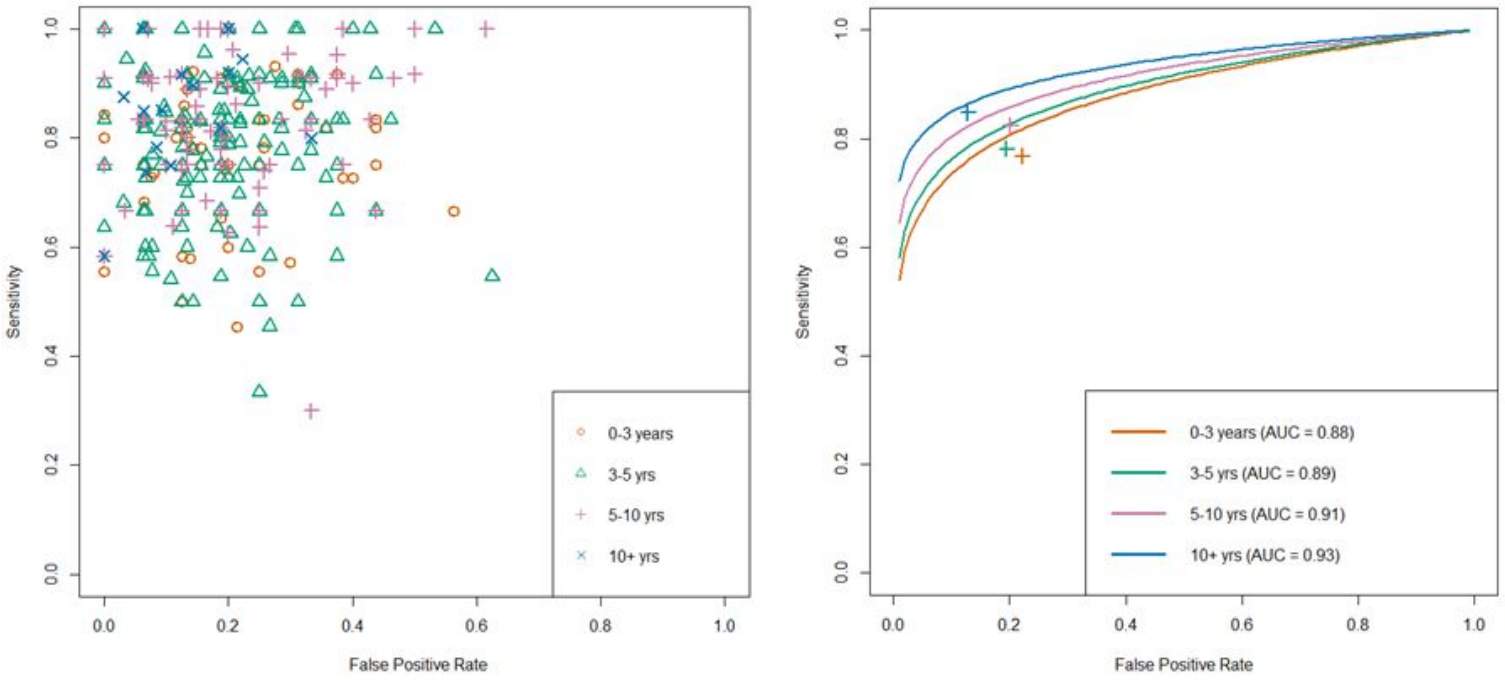


Figure 5: The individual performance of 301 human readers in Tschandl et al[18] stratified by experience level (5a, left) and summarised with SROC analysis (5b, right) using a random effects model (coloured lines) as well as the average of sensitivity and specificity (coloured crosses).

Conclusion

The estimation of average human performance is an important application of MRMC studies, both in diagnostic specialties such as radiology and in pre-clinical studies comparing human performance with that of AI models.

In the diagnostic radiology literature, ordinal scoring systems have been widely used despite the relative lack of these in clinical practice, their biological implausibility, and the readers’ lack of

experience with them. In the AI literature, average human performance has been variably reported but the most common method has been to pool sensitivity and specificity independently, a technique which is methodologically flawed and will consistently bias results in favour of the AI models.

We have described the use of well validated meta-analytic techniques for the purpose of estimating average human performance where the readers produce binary diagnostic labels, and have shown the benefits of doing so by re-evaluating a number of heavily cited medical AI papers. These results show improved estimation of performance, as well as other attractive properties including providing a single metric for discrimination performance and the ability to produce estimates of variance that incorporate both the number of cases as well as the number of readers. In at least one case (CheXNet) the latter property may have altered the interpretation of a published experiment, revealing that the reported difference between human and AI performance in that work was not compelling.

These methods are not technically novel nor are they complicated, simply involving the fitting of bivariate linear models. The value of applying epidemiological meta-analytic techniques to medical AI problems arises from the availability of extensive practical experience and methodological literature regarding these techniques, the wide availability of statistical libraries to perform these operations in most common programming languages, and the flexibility of the methods. We believe that this approach can be used to standardise assessment of reader studies with binary outcomes, improving the quality and validity of these experiments in both diagnostic medicine and medical AI research.

References

1. Carter, J. V., Pan, J., Rai, S. N. & Galandiuk, S. ROC-ing along: Evaluation and interpretation of receiver operating characteristic curves. *Surgery* **159**, 1638–1645 (2016).
2. Nagendran, M. *et al.* Artificial intelligence versus clinicians: systematic review of design, reporting standards, and claims of deep learning studies. *BMJ* **368**, m689 (2020).
3. Irwig, L. *et al.* Guidelines for meta-analyses evaluating diagnostic tests. *Ann. Intern. Med.* **120**, 667–676 (1994).
4. Gatsonis, C. & Paliwal, P. Meta-analysis of diagnostic and screening test accuracy evaluations: methodologic primer. *AJR Am. J. Roentgenol.* **187**, 271–281 (2006).

5. Macaskill, P., Gatsonis, C., Deeks, J., Harbord, R. & Takwoingi, Y. Cochrane handbook for systematic reviews of diagnostic test accuracy. *Version 0. 9. 0. London: The Cochrane Collaboration* (2010).
6. Djulbegovic, B. & Guyatt, G. H. Progress in evidence-based medicine: a quarter century on. *Lancet* **390**, 415–423 (2017).
7. Littenberg, B. & Moses, L. E. Estimating diagnostic accuracy from multiple conflicting reports: a new meta-analytic method. *Med. Decis. Making* **13**, 313–321 (1993).
8. Rutter, C. M. & Gatsonis, C. A. Regression methods for meta-analysis of diagnostic test data. *Acad. Radiol.* **2 Suppl 1**, S48–56; discussion S65–7, S70–1 pas (1995).
9. Holling, H., Böhning, W. & Böhning, D. Meta-analysis of diagnostic studies based upon SROC-curves: a mixed model approach using the Lehmann family. *Stat. Modelling* **12**, 347–375 (2012).
10. Reitsma, J. B. *et al.* Bivariate analysis of sensitivity and specificity produces informative summary measures in diagnostic reviews. *J. Clin. Epidemiol.* **58**, 982–990 (2005).
11. Dendumrongsup, T. *et al.* Multi-reader multi-case studies using the area under the receiver operator characteristic curve as a measure of diagnostic accuracy: systematic review with a focus on quality of data reporting. *PLoS One* **9**, e116018 (2014).
12. Obuchowski, N. A. & Bullen, J. A. Receiver operating characteristic (ROC) curves: review of methods with applications in diagnostic medicine. *Phys. Med. Biol.* **63**, 07TR01 (2018).
13. Gallas, B. D., Pennello, G. A. & Myers, K. J. Multireader multicase variance analysis for binary data. *J. Opt. Soc. Am. A Opt. Image Sci. Vis.* **24**, B70–80 (2007).
14. White, C. S., Pugatch, R., Koonce, T., Rust, S. W. & Dharaiya, E. Lung Nodule CAD Software as a Second Reader. *Academic Radiology* vol. 15 326–333 (2008).
15. Gallas, B. D. *et al.* Evaluating imaging and computer-aided detection and diagnosis devices

at the FDA. *Acad. Radiol.* **19**, 463–477 (2012).

16. Gur, D., Rockette, H. E. & Bandos, A. I. 'Binary' and 'Non-Binary' Detection Tasks: Are Current Performance Measures Optimal? *Acad. Radiol.* **14**, 871–876 (2007).
17. Rajpurkar, P. *et al.* Deep learning for chest radiograph diagnosis: A retrospective comparison of the CheXNeXt algorithm to practicing radiologists. *PLoS Med.* **15**, e1002686 (2018).
18. Tschandl, P. *et al.* Human–computer collaboration for skin cancer recognition. *Nature Medicine* vol. 26 1229–1234 (2020).
19. Team, R. C. & Others. R: A language and environment for statistical computing. (2013).
20. Doebler, P. & Holling, H. Meta-analysis of diagnostic accuracy with mada. *R Packag* **1**, 15 (2015).
21. Esteva, A. *et al.* Dermatologist-level classification of skin cancer with deep neural networks. *Nature* **542**, 115–118 (2017).
22. Bien, N. *et al.* Deep-learning-assisted diagnosis for knee magnetic resonance imaging: Development and retrospective validation of MRNet. *PLoS Med.* **15**, e1002699 (2018).
23. Rajpurkar, P. *et al.* CheXNet: Radiologist-Level Pneumonia Detection on Chest X-Rays with Deep Learning. *arXiv [cs.CV]* (2017).
24. McKinney, S. M. *et al.* International evaluation of an AI system for breast cancer screening. *Nature* **577**, 89–94 (2020).

2.2 From “how well does the model perform?” to “what is the worst mistake the model can make?”

The most important cause of the implementation gap in preclinical medical AI testing is not related to statistical analysis or metric choice *per se*, but instead is related to the underlying nature of AI systems. Modern AI systems learn to recognise useful features from their input data which they use to solve their task, and, unlike humans, are not limited to using features which would be considered sensible or biologically plausible. Indeed, modern AI models tend to be *underspecified*; given a model architecture and a training data set, there are many different ways in which the model can learn to solve the task, even if there is only one intended solution (i.e., “do the task in the way a human would, but better if possible.”) [37].

While humans tend to rely on features which are biologically plausible, causally linked to the task, and justified by expert opinion or even common sense, AI systems will make use of any features which are useful for the task and are learnable from the training data. This often leads to AI models which rely at least partially on unintended and unwanted solutions, such as identifying any canine on a snowy background to be a wolf, and any canine on grass to be a dog [38].

These unintended solutions can lead to unexpected or aberrant model behaviour, where the model appears to work well in most circumstances but can fail on cases which humans tend to succeed at, and most concerningly, cases where errors are potentially more harmful to patients.

Given this property of medical AI, it is not enough to ask how well AI models perform in aggregate on datasets containing a variety of cases. Instead, we need to ask “what are the worst mistakes that the model makes?” This question is grounded in clinical best practice; it requires human expertise to recognise what makes certain mistakes worse than others.

In “**The Medical algorithmic audit”** (in press) I describe the formal process of thoroughly auditing medical AI models, intended to provide a mechanism for clinical users, developers, and regulators to identify potential sources of the implementation gap, and ideally to be able to remediate these issues prior to the marketing of AI devices.

The Medical Algorithmic Audit

Publication status: In press, *Lancet Digital Health, 2021.*

Contribution: 30%

Detailed description contribution: This piece of work was complicated and multidisciplinary, including input from clinicians, machine learning experts, trialists, statisticians, and ethicists. I was responsible, along with first author Dr Xiao Liu, for the largest contribution towards this work, including the conception of the work, leading the literature search, and the planning and drafting of the manuscript.

The remaining authors Drs McCradden, Glocker and Professors Ghassemi and Denniston each were involved in editing the document and contributed an estimated 10% of the work towards the publication.

Certification from co-authors:

Xiaoxuan Liu Ben Glocker Marzyeh Ghassemi

Melissa McCradden Alastair Denniston

Declaration: This publication was part of the work undertaken during and for my HDR candidature.

The Medical Algorithmic Audit

Xiaoxuan Liu[1,2,3,4,5], Ben Glocker[6], Melissa McCradden[7,8], Marzyeh Ghassemi[9,10], Alastair K. Denniston*[1,2,4,5,11], Luke Oakden-Rayner*[12]

1. Academic Unit of Ophthalmology, Institute of Inflammation and Ageing, College of Medical and Dental Sciences, University of Birmingham, UK
2. Department of Ophthalmology, University Hospitals Birmingham NHS Foundation Trust, Birmingham, UK
3. Moorfields Eye Hospital NHS Foundation Trust, London, UK
4. Health Data Research UK, London, UK
5. Birmingham Health Partners Centre for Regulatory Science and Innovation, University of Birmingham, Birmingham, UK
6. Biomedical Image Analysis Group, Department of Computing, Imperial College London, South Kensington Campus, London, UK
7. The Hospital for Sick Children, Toronto, Canada
8. Dalla Lana School of Public Health, Toronto, Canada
9. Department of Computer Science, University of Toronto, Toronto, Ontario, Canada
10. Vector Institute, Toronto, Ontario, Canada
11. National Institute of Health Research Biomedical Research Centre for Ophthalmology, Moorfields Hospital London NHS Foundation Trust and University College London, Institute of Ophthalmology, London, UK
12. Australian Institute for Machine Learning, University of Adelaide, Adelaide, Australia

*Joint corresponding authors; a.denniston@bham.ac.uk, luke.oakden-rayner@adelaide.edu.au

Degrees:

Xiaoxuan Liu: MBChB

Ben Glocker: PhD

Melissa McCradden: PhD

Marzyeh Ghassemi: PhD

Alastair K. Denniston: PhD (full professor)

Luke Oakden-Rayner: MBBS

Institutional correspondance:

Luke Oakden-Rayner: The Australian Institute for Machine Learning, Lot 14, Corner Frome Road and, North Terrace, Adelaide SA 5000. Ph: +61 (08) 8313 3051.

Alastair K. Denniston: Institute of Inflammation and Ageing, College of Medical and Dental Sciences, University of Birmingham, Edgbaston, Birmingham, B15 2TT, UK

Summary

Artificial intelligence systems for healthcare, like any other medical device, have the potential to fail. However, specific qualities of AI systems, such as the tendency to learn spurious correlates in training data, poor generalisability to new deployment settings, as well as lack of reliable explainability mechanisms, mean they can yield unpredictable errors which may be entirely missed in the absence of proactive investigation.

In this paper, we propose a medical algorithmic audit framework that guides the auditor through a process of considering potential algorithmic errors in the context of a clinical task, mapping the components which may contribute to the occurrence of errors, and predicting their potential consequences. We suggest several approaches for testing algorithmic errors including exploratory error analysis, subgroup testing and adversarial testing, and provide illustrative examples from our own work and from previous published papers.

The algorithmic audit is a tool which can be used to better understand the weaknesses of an AI system and put in place mechanisms to mitigate their impact. Importantly, we propose that safety monitoring and algorithmic auditing should be a joint responsibility between users and developers, and encourage the utilisation of feedback mechanisms between them to promote learning and maintain the safe deployment of AI systems.

Keywords: artificial intelligence, machine learning, deep learning, algorithmic audit, clinical audit, algorithmic error, safety, algorithmovigilance, failure modes and effects analysis

Introduction

Advances in artificial intelligence (AI) have attracted significant interest for their potential applications in healthcare, particularly systems based on deep learning and neural networks. A vast body of literature has been published proposing AI/machine learning-based solutions for disease detection, classification, prediction, or even as therapeutic interventions including titration of drug dosages or offering mental health support through AI 'chatbots'.(1,2)
More recently, there has been a shift in emphasis from reporting impressive performance results to actively investigating algorithmic reliability and characterising algorithmic errors.(3–6) Indeed, the analysis of algorithmic errors is a new minimum reporting requirement in the recently published SPIRIT-AI (Standard Protocol Items: Recommendations for Interventional Trials–Artificial Intelligence) and CONSORT-AI (Consolidated Standards of Reporting Trials–Artificial Intelligence) guidelines for reporting clinical trial protocols and reports for AI interventions.(7,8) This change in focus from *'what is the best performance this AI system can achieve*' to '*what is the worst mistake this AI system can make*' aligns with the foundational maxim embedded in medical safety: to first do no harm. Traced back to Hippocrates, the clinician's commitment to abstain from and minimize harm is consecrated in modern medical ethics as the principle of non-maleficence, which recognises that medical interventions carry inherent risks which must be mitigated. The question of medical AI safety is being asked at a crucial time, where this is no longer a theoretical concern but one of preventing harm both for the AI systems that have already received regulatory approval and for the many more that are in various stages of development but which are intended for clinical deployment.

So why are AI systems different? Concerns have been raised that, unlike other medical interventions, AI systems can yield errors that are difficult to foresee or prevent, due to the very nature of these systems. Modern AI systems based on deep learning establish complex and opaque mathematical relationships between the input data and the output predictions, with little to no human control over how those predictions are generated. While this gives rise to a powerful machinery for learning patterns in the data, there is also a significant risk for the machine to pick up spurious correlations; relationships which appear useful in the training context but are unreliable when applied to real-world patients. An example might be an AI model

that learns to detect surgical skin markings to diagnose skin cancer, rather than looking for features related to the lesion itself(9). Importantly, the errors of AI systems appear to be quite distinct from the errors of human experts. In medical imaging for example, the majority of human errors (60-70%) are related to perceptual failure, caused by factors such as the subtlety of visual findings, incomplete search of the entire image, and “satisfaction syndrome” (where finding an abnormality makes the reader less likely to find a second one)(10). In contrast, AI cannot fall victim to incomplete search or satisfaction syndrome. In this context, it is entirely reasonable to expect that AI systems of equal performance to human readers will produce *different* errors, which can lead to different clinical outcomes.

The "AI performance gap" can be caused by a variety of factors, including those related to the algorithm development, the input data used during training and deployment and interactions with users and the deployment environment. From the algorithm development, the model design, training strategies, as well as the choice of training data (in the case of poorly labelled or under-representative data) can directly influence the AI performance. Mismatch/incompatibility of input data used during deployment can arise from various types of dataset shift, as described by Castro et al (including population shift, annotation shift, prevalence shift, manifestation shift and acquisition shift)(11). Interactions with users and the deployment environment are subject to automation bias, human error and unintended/intended misuse of the AI system(12). It is also worth noting that the reasons for unexpectedly poor clinical performance can be non-obvious even after human inspection, and subtle or even unnoticeable differences in the input data may lead to catastrophic failure. This relates to the underlying mathematical approximations that AI systems use to map input data (e.g., a medical scan) to target outputs (e.g., a diagnostic label). Generally, we can assume that AI systems will operate well within the space mapped out by the training data (a process called “interpolation”), but perform rather poorly on out-of-distribution data which requires extrapolation. Intuitively, the further an input sample is away from the statistical distribution of the training data, the more unpredictable becomes the behavior and outputs of the AI system. Unfortunately, given the complexity of most forms of medical data, it can be difficult to define which cases are in-distribution and which are out-of-distribution. Furthermore, this drop in performance may not be obvious at the aggregate level of typical AI testing, but rather in subsets of the target cohort or specific strata represented within the input data: a concept which has been described as ‘hidden stratification’.(13) These factors all

contribute to the performance gap between preclinical testing and real-world deployment, and current evaluation strategies are ill-suited to identifying the problem.(14)

What is an algorithmic error and why do we need to know about them?

We define algorithmic errors as any outputs of the AI system which are inaccurate, including those which are inconsistent with the expected performance and those which can result in harm if undetected or detected too late. Within these, there is a category where the output may be correct but the algorithm is clearly informed by a flawed decision-making process. We suggest that these are also considered algorithmic errors, as this indicates high risk of future errors and should therefore be treated with similar levels of caution. An example might be when a skin-cancer detection AI system determines a lesion is malignant because of the presence of surgical skin markings rather than relying on the visual characteristics of the lesion.[9] Where there is a pattern or systematic nature to the occurrence of errors, we refer to this as a failure mode: the tendency to malfunction in the presence of certain conditions. Whereas an error can be a single occurrence, failure modes represent errors which will repeatedly occur and often have similar consequences. Whilst individual errors may not always result in direct harm, their frequency or the summation of multiple errors may reach above an acceptable threshold and result in overall harm.

By proactively investigating algorithmic errors and failure modes, the auditor becomes better placed to monitor AI systems effectively and to understand the potential failure modes and their consequences. Within the broader mandate to ensure AI systems are safe, undertaking regular systematic analyses of the observed errors is helpful for a number of reasons:

1. It is an essential component of safety monitoring and adverse event reporting.(15,16)
2. It allows quantification of risk for the AI system, which can be weighed against the potential benefits, to inform decision-making around whether it is appropriate to apply the model clinically. There may be existing benchmarks within current clinical practice (such as estimated human radiologist error rates for a diagnostic task) which would inform the risk-benefit-ratio for deploying the AI system.
3. It may reveal unknown failure modes of the AI system, such as tendency to produce higher error rates in certain populations, diseases or settings, or in the presence of specific input data characteristics.(9,11,17)

4. Prior to deployment it can be used to derive a measurable adverse event rate, which can inform how closely safety monitoring and post-deployment auditing should be performed. It also provides a baseline measurement against which ongoing performance can be benchmarked.
5. It can inform risk mitigation strategies so that those overseeing deployment of the AI system can anticipate errors if the conditions known to trigger failure do occur, put in place measures to avoid failure modes and preemptively set ‘hard stops’ in high risk situations.
6. It can provide valuable feedback and information for future AI development and model improvement, and also highlight potential need for post-deployment calibration and/or localisation of AI systems.
7. It can reveal systematic differences in performance across features mapping onto a protected identity or social determinant. Insight into these performance differences can prevent a systematic disadvantage to those groups resulting from the implementation of the algorithm.

In this paper, we propose an audit-based approach for investigating algorithmic errors. Algorithmic audit broadly focuses on development processes and embedding organisational principles and values in the algorithm design, and these values can vary widely depending on the organisation and context of the deployment. In the context of medical AI the audit process is more tightly focused on the safety and quality of medical systems, the outcomes and perceptions of the patient and the public, the responsible utilisation of healthcare resources, and the equitable distribution of healthcare and healthcare outcomes.

The Principles Underpinning the Medical Algorithmic Audit

The importance of safety and quality for medical algorithms is embedded in the principles of medical ethics which describe the obligations of clinicians to their patients and the public. Evidence-based practice reflects the ethical imperative to act to promote the patient’s best interests (beneficence) while minimizing harm (non-maleficence), with empirical data forming part of the foundation upon which these judgments are made in consort with patient values. Typically, the information gathered through the process of prospective evaluation is contextualized to a clinical setting on the basis of factors relating to each individual patient(18).

For interventions like drugs, the intervention itself is identical for all iterations (e.g., the chemical structure of a single pharmaceutical agent is the same for every patient who takes it) and it is within individuals that responses vary. With AI systems, the intervention is acutely sensitive to between-individual and within-individual feature variation, as the very power of the computational technique is in its ability to utilise feature variations to make individual-level predictions. However, what AI systems cannot do is apply clinical knowledge and domain expertise (including prior experience, contextual understanding and causal knowledge) 'common sense' to distinguish between relevant feature variation due to disease versus irrelevant feature variation due to other biological confounders or non-biological sources, potentially resulting in unreliable predictions. This means that to translate algorithms within clinical practice, more nuanced information is required describing the algorithm's performance across a range of relevant features, which is the goal of medical auditing. This information then forms the constellation of knowledge and practices that guides effective - and beneficial - translation of interventions(18).

A core and often overlooked concern with AI is that of fairness. So long as bias and social determinants of health exist, these patterns will entrench themselves within healthcare ML. In many cases the performance of AI models differs across patient identities or social determinants of health (often proxies for identities), which can pose a threat to another core ethical principle: justice. In this case, we might consider distributive justice as a desirable property of AI-enabled care delivery (i.e. whether the benefits afforded by ML are conferred equally to all). Distributive justice also points us to the necessity of redressing disparities. If an audit reveals substandard performance among certain groups, compensatory mechanisms may help ensure these patients are not disadvantaged by use of the algorithm. Medical auditing can reveal areas where these mechanisms are required and also point to how potential disadvantages may be redressed.

The elements of a medical algorithmic audit

In this paper we build on the algorithmic audit approach proposed by Raji et al(19), who describe a qualitative structured audit process applying the "SMACTR" framework (Scoping, Mapping, Artifact Collection, Testing, and Reflection) to AI, as a general purpose technology. Whilst this framework was originally proposed as a way of assessing whether AI development

was conducted in alignment with the principles of an organisation, its structure is highly applicable to local auditing of AI performance due to its orientation towards *internal* auditing (and thus led by those closest to implementation). Each step of the SMACTR framework has its own set of documentation requirements, thus facilitating accountability and iterative, ongoing safety monitoring. There is also emphasis on other established auditing practices in medicine and other industries, including process mapping, failure modes and effects analysis, risk prioritisation and planning mitigating actions. We adapt this framework for use in medical AI applications **(Figure 1)** and approach the problem from two perspectives: that of the developer, who has the ability to modify the AI system in response to the audit results; and that of the user, who cannot modify the AI system but has means to set up risk mitigation plans specific to the deployment setting. We apply the principles of failure modes and effects analysis (the FMEA tool), a well established mechanism in engineering to facilitate risk assessment, risk prioritisation and risk mitigation. For illustrative purposes, an example of an audit for a hip fracture detection algorithm is published as supplementary information in Oakden-Rayner *et al* (included in thesis) alongside a detailed breakdown of the FMEA. The benefits of performing the FMEA is to initiate and guide a critical thought process, rather than to establish whether the AI system is acceptable or unacceptable or to provide certainty that all risks can be anticipated and minimised. FMEA has previously been applied to clinical settings, although must be interpreted with care due to issues around reproducibility and incompleteness(20).

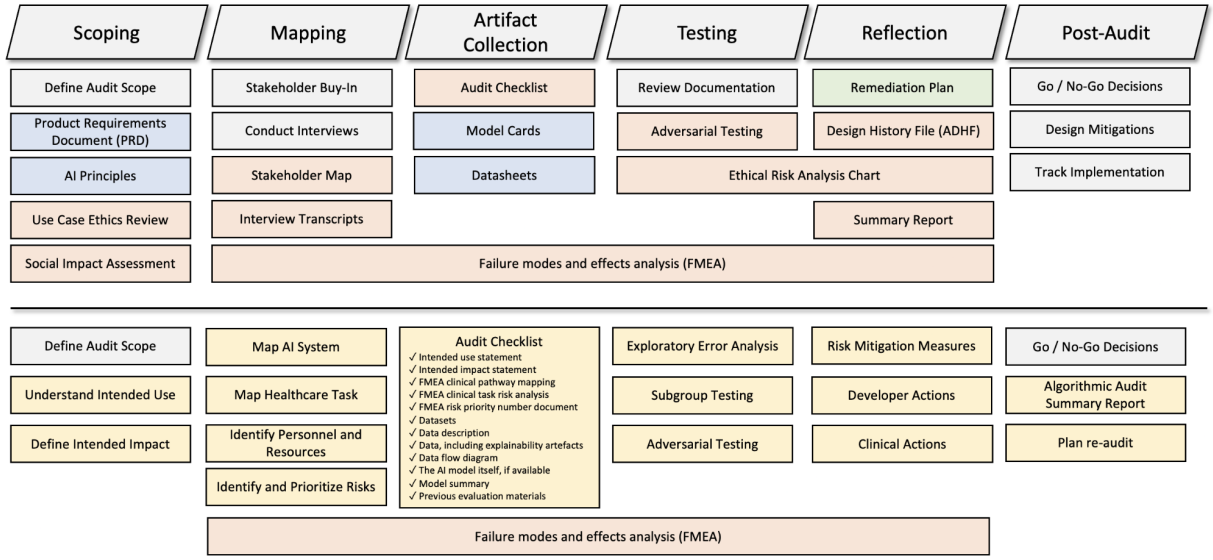


Figure 1: **Top**: Overview of the Internal Audit Framework from Raji *et al(20)*. Gray = a process, and colored sections represent documents. Orange = produced by the auditors, blue = produced by the engineering and product teams, and green = jointly developed. Image reproduced with permission. **Bottom**: Proposed modifications for the Medical Algorithmic Audit.

The medical algorithmic audit may be conducted by AI developers but is also likely to be conducted by stakeholders with no involvement in algorithm design, such as healthcare workers. It is possible that during deployment, a myriad of human factors combined with a poor understanding of AI systems may create a situation where all errors are assumed to be a fault of the algorithm's design. It is therefore essential that clinical auditors have the necessary tools to identify error sources which are preventable (input data factors, user factors) and not preventable (factors which are intrinsic to the algorithm itself). Whilst those outside of the development team may have no opportunity to change the algorithm, they may be able to control or influence the circumstances under which it is deployed, which is intrinsically tied to the likelihood of errors, as well as the ability to avoid or manage them. To consider the medical algorithmic audit from both perspectives, **Table 1** describes tasks taken by users and developers separately for each stage of the audit. It is also worth noting that developer and clinical audits are complementary; ideally both will be performed either separately or in collaboration.

Table 1. Checklist for artefact collection

	Developer actions	User actions
Scoping	*Define* intended use	*Identify* intended use
	Anticipate intended impact(s)	*Define* intended impact(s)
Mapping	Mapping of the AI system	Mapping of the AI system
	Define data flow	Define data flow
	Identify known risks of the AI system: - Existing published evidence - Existing unpublished evidence - Through knowledge of the training data	Identify known risks of the AI system: - Existing published evidence
		Identify known risks of the *healthcare task*
		Mapping of the healthcare task including elements before and after the AI system in the clinical pathway - Identify important patient or data subgroups. - Identify potential sources of atypical input data. - Identify relevant outcomes to be measured and how they will be captured in the audit.
	Summarise risks in a risk priority number	Summarise risks in a risk priority number
Artifact Collection	Intended use statement	Intended use statement
	Intended impact statement	Intended impact statement
	FMEA: - Clinical pathway mapping - Clinical task risk analysis - Risk priority number document	FMEA: - Clinical pathway mapping - Clinical task risk analysis - Risk priority number document
	Datasheet for datasets (training and test data)	Datasheet for datasets (deployment data)
	Data flow diagram	Data flow diagram
	The AI system itself	The AI system itself
	Model Summary	Model Summary
	Data for direct assessment, including explainability artifacts and adversarial testing artifacts	Data for direct assessment, including explainability artifacts and adversarial testing artifacts
	Previous evaluation materials (including performance testing/user experience artifacts)	Previous evaluation materials (including performance testing/user experience artifacts)

Testing	Exploratory error analysis - False positives and false negatives - Explainability methods (saliency maps and feature weights)	Exploratory error analysis - False positives and false negatives - Explainability methods (saliency maps and feature weights)
	Subgroup testing - Table 1 subgroup analysis - Task-specific subgroup analysis	Subgroup testing - Table 1 subgroup analysis - Task-specific subgroup analysis
	Adversarial testing	Adversarial testing (if possible)
Reflection	Risk mitigation measures: - Retrain the model - Modify the model threshold - Modify the workflow or intended use	Risk mitigation actions - Continue use with additional human oversight - Modify or limit use - Withdraw use
	Compile algorithm audit summary report and share with relevant stakeholders.	Compile algorithm audit summary report and share with relevant stakeholders.

Scoping

Scoping is the process of defining the intended purpose of the AI system and anticipating potential harms that may arise due to use of the system. In Raji *et al*, the framework is intended for any domain where AI may be applied. In the setting of medical AI testing, the scope of the audit is more clearly defined: the ethical and clinical motivation is uniform across medical AI studies, with the intention to improve healthcare outcomes (i.e., wellbeing, financial, organisational) and to promote distributive justice. Therefore, scoping in medical algorithmic audit should focus on two key elements, the intended use, and the intended impact.

The **intended use** is a term from regulatory guidance (21,22) which describes a high level summary of how the device is to be applied. The United States Food and Drug Administration (FDA) Premarket Approval (PMA) guidance states "*Indications for use for a device include a general description of the disease or condition the device will diagnose, treat, prevent, cure, or mitigate, including a description of the patient population for which the device is intended. Any differences related to gender, race/ethnicity, etc. should be included in the labeling*." The

intended use statement is defined by the developer, who has knowledge of any prior evidence supporting indications for legal and safe use. It should also be known to the user, who decides whether the intended use statement matches the clinical task and clinical pathway in which the algorithm is intended to be deployed. For example, in the hip fracture audit (included in thesis), scoping of the intended use refers to the function of the algorithm (detecting proximal femoral fractures) as well as its integration into a clinical pathway (where detection leads to admitting under an orthopaedic team and booking of further imaging if necessary). Other considerations include any limits on the healthcare environment for use (i.e., inpatient, outpatient) and the intended users/oversight (i.e., health professional, patient, autonomous). It is important to establish a clear understanding of whether the current application falls within the AI system's intended use, or if there are areas of ambiguity (i.e., from missing or poorly defined intended use descriptions). Any identified mismatches can motivate a targeted error analysis during the algorithmic audit.

The **intended impact** identifies the clinical or healthcare target of the AI system, accompanied by the ensemble of information that describes the boundaries within which the system is efficacious(18). This high-level statement describes how the AI system will affect healthcare outcomes if it works as intended. The developer may be able to define, in theory, the intended impact, but the user is better placed to consider this. The hip fracture audit (included in thesis) has several intended impacts including reduced time to admission and theatre, reduced resource utilisation in the emergency department and unnecessary imaging, and downstream improvement in health outcomes.

Different users of the same algorithm may have different target impacts specific to their health setting and needs. They may choose to implement them in different ways to produce different results and therefore their measures of success (and failures) will also be different. The auditor should consider whether there are any unacceptable high risk outcomes or adverse events (such events in medical safety are distinct, as they are considered so severe that they should never occur, such as "never events" like surgical procedures performed on the wrong limb). It may be helpful to consider this also in the context of non-AI systems with the same or similar intended use and intended impacts.

Both the intended use statement and intended impact statement will be used during the next phase (mapping), as these statements define the scope of algorithmic errors related to use of the AI system.

Mapping

The mapping phase considers two main topics; the mapping of personnel and resources necessary for the audit, and the mapping of the risks and known vulnerabilities of the intended use as the first stage of the Failure Modes and Effects Analysis.

Personnel who may be useful for medical algorithmic audit include developers, users, and domain experts (including medical ethicists), particularly those who have experience with the AI system. Developers have a substantial role to play in terms of providing periodic evaluations to guarantee expected performance, as is the case with other medical devices such as scanners, which often include 24/7 service and support plans to ensure the device continues to meet operational, regulatory, quality and safety requirements. Where possible, developers should also design mechanisms which allow the user to carry out audits independently at a local level. Resources that may be required include, but are not limited to, access to suitable training and/or testing data and the associated labels (including non-target labels such as demographic information and hospital process factors), access to model predictions on the test data, access to any interpretability tools produced for use with the AI model, and access to the model itself if a more in-depth introspection or further data challenges (such as adversarial testing) may be required.

Failure Modes and Effects Analysis (FMEA) is a prospective risk analysis tool which first maps out the process, and then is used to identify foreseeable failures which may occur. In the mapping phase there are two important elements of FMEA: mapping of the AI system itself and the mapping of the healthcare task.

Mapping of the AI system itself is a detailed expansion on the intended use statement and analysis of prior evidence documenting risks intrinsic to the AI system by design, or which the AI system has encountered previously. This may include an evaluation of the existing literature on known risks, or a scoping of other AI systems with similar intended use for potential risks.

Mapping the AI system also involves mapping any prerequisite steps or minimal requirements which are essential to achieving expected performance. Crucial to this is the process for handling and selection of input data, which is sometimes underspecified and poorly reported.

Mapping of the healthcare task is a contextualised analysis of the AI system as a component of clinical care within a health system. It requires clinical knowledge of the use-case, the clinical workflow (including existing safeguards for detecting errors), user behaviour (healthcare provider, patient and public) and understanding of the potential consequences of errors. This knowledge can complement the mapping of the AI system above, to anticipate when, and how, failure modes can arise. Mapping the clinical pathway can identify upstream factors which may increase the chances of algorithmic error, and downstream consequences which may occur as a result of algorithmic error. It also involves identifying any important patient or data subgroups and any specific features of the input data which are unusual or atypical.

It can be helpful at this stage to map the AI system in relation to the clinical task and intended impacts in the form of a causal diagram, to determine the direction of causality between variables measured in the audit(23). This will inform the metadata required for the *Artifact Collection* phase and can help auditors in making sense of relationships and assumptions between relevant components of the healthcare task in the *Reflection* phase.

These elements are then summarised in a risk priority number (RPN), which ranks the identified risks. The risk priority number is an arbitrary value calculated through ranking and combining three elements: severity (severity of the failure effects), occurrence (likelihood of occurrence) and detection (effectiveness of mechanisms to detect the failure before it results in adverse consequences). The ranking of each item is subjective but a scale should be defined so that RPNs in future audits can be comparable. For example, in the hip fracture audit (included in thesis), severity and occurrence were scored between 1 to 4. A severity score 4 was catastrophic (failure could cause injury or death, extreme loss of trust) and 1 was minor (no intervention needed, no injury to patient). An occurrence score of 4 was frequent (several times in 1 day) and 1 was remote (may happen in more than 6 months). The auditor should decide whether all three elements are applicable to the AI system, or whether there are additional elements which should be added. In the hip fracture audit only severity and occurrence elements were included as most

risks could not be easily detected in the current clinical workflow (resulting in homogenous scoring).

It is critical to understand that the actual RPN value is not a measure of safety, nor should there be an attempt to create arbitrary thresholds to determine the acceptability of risks. Rather, the RPN enables relative ranking of all risks to prioritise those which need urgent attention and to serve as a baseline for re-evaluation in future audits.

Artifact collection

The *artifact collection phase* involves gathering the documents and materials identified in the *mapping phase* which may inform the audit **(Table 1)**. There are three main components to consider in the context of medical AI systems (aside from those already produced in the scoping and mapping phases): 1) any relevant datasets (training data, previous evaluation data and/or prospectively collected audit data for the current audit), 2) the model itself, and 3) results of previous evaluations of the model or task.

The **datasets** are of primary importance in determining both the performance of the AI system, and the potential limitations and failure modes. Various datasets come into play throughout the development, evaluation and monitoring of AI systems and all are relevant for the algorithmic audit, but may reveal different information about errors and failure modes. The relevant datasets are the algorithm training data (for developing the algorithm, which may include data for internal validation), previous test data (for evaluation or validation of the algorithm *in silico*) and deployment data (data generated as a byproduct of the algorithm being used). Both the test data and deployment data can be used in an algorithmic audit, however the information provided within them may vary. Note that evaluation data, and in particular *labelled* evaluation data can be difficult or impossible to obtain in live deployment situations (or in certain evaluation designs, such as RCTs of effectiveness), where the ground truth for each case is not routinely collected. In these settings the identification of sources of weak labels (such as adverse events registers and user feedback) will be important, and the limitations of these labels should be clearly indicated.

There is often little relationship between errors on the training set and errors that occur during deployment, therefore access to the complete training data is a lower priority in algorithmic audit. Although training data may be used for conducting exploratory error analyses (discussed below), it is not uncommon for deep learning models in particular to achieve negligible training error but still perform poorly in a test or deployment environment. Access to this data can also be problematic for users and external auditors, given the commercial value of this data and the sheer size of these datasets.

Whilst direct access to the relevant data is likely to be useful, understanding the data processes is equally important. This information can be formalised with a “datasheet”(24) and a data flow diagram. A datasheet provides an extensive description of the data generating process, dataset collection, dataset composition, and dataset processing and labelling. Datasheets can be extremely valuable during an audit, as the dataset composition (in particular, the training data composition) can suggest likely failure modes (for example, patient subgroups that are under-represented in the training data). Access to datasheets is not similarly problematic or commercially sensitive, and should be provided by the developers wherever possible. In addition to the datasheet, there should be a data flow diagram which outlines the handling of data from point of acquisition to presentation to the algorithm. This should include any pre-processing steps, such as data transformation and normalisation, as well as exclusions based on data quality and a traceability mechanism for unusable or discarded data. Any results from previous explainability methods such as saliency or attention maps, per case feature importance measures, feature visualisations and so on should also be collected at this stage.

The **model** itself is also important in the audit process. Basic information about the model design, version, and model developers should be collected as a minimum. Such information can be summarised in a “model card”.(25) If the AI system consists of multiple components (for example a segmentation step, followed by a classification step(26)), artifacts should be collected for each individual component where possible. Where multiple audits have been conducted over time spanning updated versions of the AI model, documentation regarding changes between updates and any published evaluations since the last audit should also be collected. While model description can be formalised with a “model card”(25), there is significant overlap between this and other artefacts to be collected.

While direct access to the model code and parameters (sometimes called a “white box audit”) can be hypothetically useful, for example by performing stress-testing of the AI system by intentionally modifying input data to induce errors, this is rarely possible due to intellectual property concerns. The majority of the benefit that access would provide can be equally obtained with the ability to test the model on new cases and receive model outputs, usually via a web portal or API (also known as a “black box audit”). Developers should provide such a mechanism for users to perform independent local testing using representative data samples, to ensure performance is as expected.

The **evaluations** performed previously are extremely important during the preparation of an algorithmic audit. Typically medical AI development goes through several phases of evaluation, and artefacts of this process include internal and external evaluation summaries, published materials on pre-clinical and proof-of-concept testing, and summaries of any previous qualitative assessments or audits. The latter may include developer and user experience materials, such as interviews, surveys, or other forms of feedback.

In the context of the hip fracture audit, the components of the scoping and mapping phases were all collated, but in addition the auditor secured access to the validation and test datasets with explainability artefacts/saliency maps for these cases, the hip fracture model itself, and documents related to model development and previous testing (27,28), including design documents for each component of the algorithm.

Testing

The most important part of the audit process, other than the implementation of recommendations, is the *testing phase*. It is also the hardest part of the process to standardise, as each AI system will face different risks and challenges and much of the assessment is informed by the results. Institutions are accountable for the choice to incorporate a given AI system into their clinical workflow, which necessitates the need to ensure its appropriateness and functionality for the particular patient populations they serve. Should an algorithm not perform as expected or if harm were to occur, an audit would provide a clear mechanism of demonstrating institutional accountability.

We suggest several key components of testing of medical AI systems during algorithmic audit:

Exploratory error analysis (EEA)

The auditors will review each example of algorithm error which has been provided (either from previous evaluations, or from detected errors/adverse events in deployment). Auditors will systematically examine both false positive and false negative groups in the case of classification systems, or outliers with high numerical errors in the case of regression models. The intent of this process is to identify any common elements among the errors (i.e., specific *types* of cases which may be more prone to error and therefore carry higher risk, as shown in the hip fracture detection example in **Figure 2**), as well as any examples of *surprising* errors (for example, a fracture detection model missing an extremely obvious fracture). Given the contrastive nature of this method, access to cases correctly analysed by the AI system can also be useful.

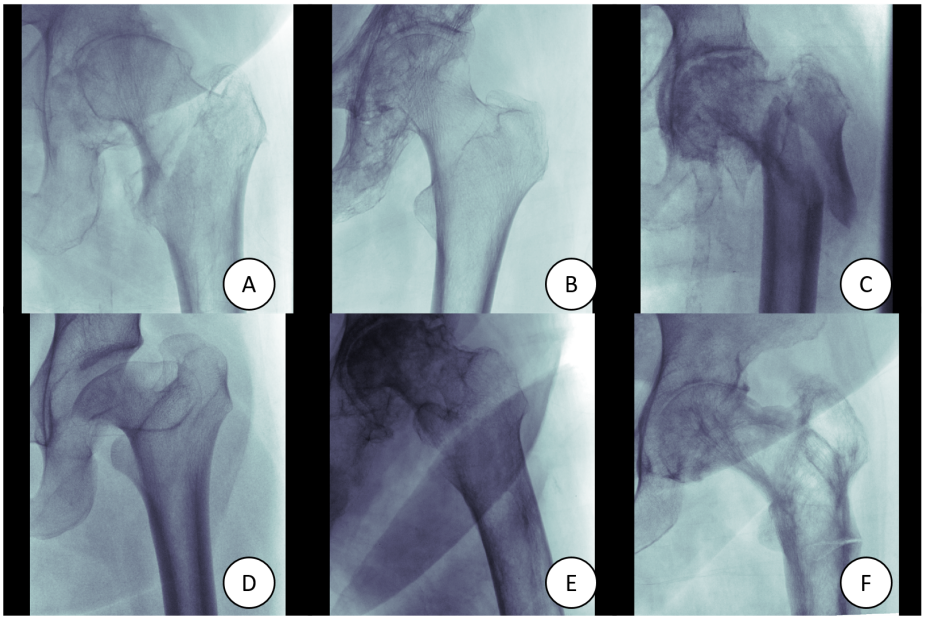


Figure 2: Audit of the hip fracture detection system (included in thesis) revealed that cases with abnormal bones or joints (Paget's disease and femoral head deformity) were overrepresented amongst the errors. The overall error rate was 2.5%, but the error rate for this subset was 50% (false positives = D, false negatives = E, F).

As EEA is exploratory in nature, it can be useful to have access to additional tools which may require access to the algorithm itself or support from the algorithm developers. Examples of

useful tools include AI explainability methods such as saliency maps and feature visualisations for image data, attention maps, feature weights or importance measures for text and tabular data, and so on. Similarly, data clustering methods have been shown to be of some help for some audit tasks such as cryptic subset detection(29). An example for the use of data visualisation is shown on the left in **Figure 3**, where, in the absence of data normalisation, the largest modes of variation in brain magnetic resonance imaging data after principal component analysis on the input images is between hospital sites. There is a high risk that a disease classification model trained on such data may pick up features associated with the site rather than the pathology, in particular if one site contributes more cases than controls. A careful data normalisation pipeline may mitigate such site differences, as shown on the right in **Figure 3**. A model trained on the normalised data may be more robust when employed to new data. While these exploratory tools are not powerful for risk assessment in isolation, they can be extremely useful during EEA.

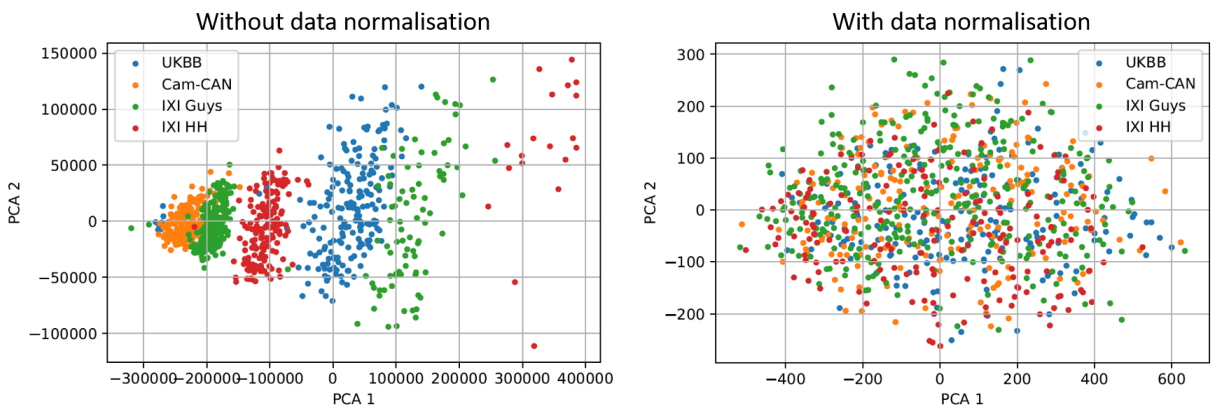


Figure 3. Principal component analysis of brain MRI, with and without data normalisation, across four hospital sites. UKBB: UK Biobank, Cam-CAN: Cambridge Centre for Ageing and Neuroscience dataset, IXI: Information eXtraction from Images dataset including data from Guy's Hospital and Hammersmith Hospital. UK Biobank data is accessed under Application Number 12579.

Subgroup testing

Subgroup testing, or secondary performance analysis, is widely used in medical and epidemiological research to investigate the possibility of confounding/stratification; patient or

data variables which indicate a subset of cases in which performance will significantly differ than in the overall cohort.

Importantly, subgroup analysis is not performed to test hypotheses. Given the reduction in power due to the lower sample sizes in subsets, as well as the inflated type 1 error rate (false positives) caused by multiple testing, these results should not be considered reliable or definitive in the same way a primary analysis may be. Instead, the goal is to identify possible high risk sub-populations within the target group. Whilst the subgroup analyses can be useful for identifying possible error patterns, these findings should be confirmed through investigation in a sufficiently powered sample.

There are three main forms of subset testing: 1) Table 1 subgroup analysis, 2) Task-specific subgroup analysis, and 3) EEA-discovered subgroup analysis.

1. Table 1 subgroup analysis

In medical publications, the baseline characteristics table - Table 1 (and sometimes Table 2) - are used to describe important forms of variation in the dataset which may cause confounding relationships within the data, or have wider implications on the results. Almost all studies report subgroups by age, sex, ethnicity, socioeconomic status, disease severity and comorbidities, and some studies also include data acquisition details such as imaging protocols and hardware devices. The reason these variables are highlighted during audit is twofold: 1) they have well known stratifying relationships with disease and treatment outcomes, and 2) if they are included in Table 1, the data is readily available and the subgroup analysis is trivial to perform. Many studies already report this information(30–32) and an example is shown from the Ting *et al* evaluation of a retinal imaging AI system **(Table 2)**.

Table 2. Example of a "Table 1 subgroup analysis" by dataset source/setting from Ting *et al*.[28] which demonstrates the performance of their AI model stratified by the clinical origin of the data.

	Table 1. Summary of External Validation Datasets for Diabetic Retinopathy							Table 5. Diagnostic Performance of AI system in External Validation		
	Nonreferable Eyes		Referable Eyes					Performance		
	No Diabetic Retinopathy	Mild Nonproliferative Diabetic Retinopathy	Moderate Nonproliferative Diabetic Retinopathy	Severe Nonproliferative Diabetic Retinopathy	Proliferative Diabetic Retinopathy	Diabetic Macular Edema	Ungradable	AUC (95% CI)	Sensitivity (95% CI)	Specificity (95% CI)
Community-based										
Guangdong	5665	1235	737	0	154	0	108	0.949 (0.943-0.955)	98.7 (97.7-99.3)	81.6 (80.7-82.5)
Population-based										
Singapore Malay Eye Study	1143	215	113	18	9	53	28	0.889 (0.863-0.908)	97.1 (95.1-99.9)	73.3 (70.9-75.5)
Singapore Indian Eye Study	1639	422	125	5	17	71	48	0.917 (0.899-0.933)	99.3 (95.1-99.9)	73.3 (70.9-75.5)
Singapore Chinese Eye Study	759	131	60	1	7	17	10	0.919 (0.900-0.942)	100 (92.5-100.0)	76.3 (72.7-79.6)
Beijing Eye Study	493	4	11	4	0	12	2	0.929 (0.903-0.955)	94.4 (72.7-99.9)	88.5 (85.4-91.2)
African American Eye Disease Study	807	50	37	5	16	28	41	0.980 (0.971-0.989)	98.8 (93.5-100.0)	86.5 (84.1-88.7)
Clinic-based										
Royal Victoria Eye and Ear Hospital	432	121	159	123	191	249	125	0.984 (0.972-0.991)	98.9 (97.5-99.6)	92.2 (89.5-94.3)
Mexican	38	284	192	51	18	223	3	0.950 (0.934-0.966)	91.8 (88.4-94.4)	84.8 (80.4-88.5)
Chinese University of Hong Kong	224	114	235	43	11	96	0	0.948 (0.921-0.972)	99.3 (97.3-99.8)	83.1 (77.9-87.3)
University of Hong Kong	1984	1485	155	14	0	214	1	0.964 (0.958-0.970)	100 (99.0-100)	81.3 (80.0-82.6)

2. Task-specific subgroup analysis (TSSA)

There is a near infinite set of possible confounding and stratifying factors in medical AI evaluation. The TSSA seeks to analyse the most concerning of these factors, and is informed by the FMEA risk analysis and risk priority score. Like the Table 1 subgroup analysis, the subgroups in the TSSA are defined prospectively, based on an understanding of the clinical task, often informed by domain experts. The main difference between a Table 1 subgroup analysis and TSSA is that the subgroups are often cryptic (unlabelled) in TSSA. It is often necessary to undertake additional labelling to identify data which is part of the subgroup of interest. As it may require significant time and resources to undertake labelling of relevant data, the risk prioritisation performed in the FMEA may inform which additional labelling should be prioritised.

Examples of task-specific factors that may be considered in TSSA include collision groups (such as a combination of features from the Table 1 analysis) and process variables such as the scanner used to obtain medical images or the presence of artefacts of medical care within the data (such as a chest drain on a chest xray for a patient being treated for pneumothorax(13) **(Figure 4)**, or a surgical mark on the skin of a patient suspected of melanoma(9)). Special

consideration should be made of data subgroups which would not be captured in a typical Table 1 analysis, for example visually distinct subsets in a medical image analysis task (such as subsolid versus solid lung nodules)(33). The TSSA may also be informed by important clinical implications associated with certain subgroups, such as the need for diagnostic certainty when differentiating infectious from non-infectious skin lesions (as the treatment for the latter, topical steroids, will often worsen the former and may make subsequent diagnosis more difficult).(34)

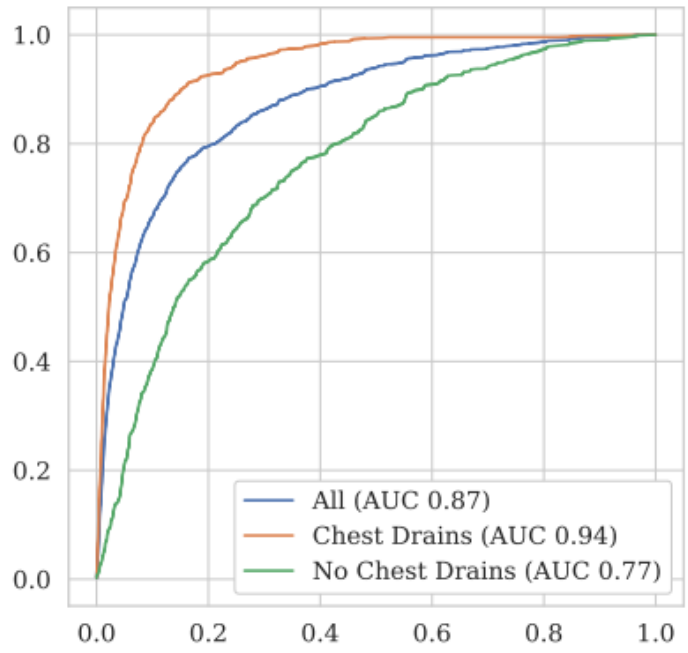


Figure 4. Example of task-specific subgroup analysis for a model detecting pneumothorax on chest radiographs, where the AI model learns to detect the artefacts of clinical care (chest drains) and fails to adequately learn the features of the pathology itself.(13)

3. EEA-discovered subgroup analysis

During the EEA process, distinct subgroups of error cases or error features may be discovered, which are not considered during Table 1 subgroup analysis or TSSA. Notably, while TSSA subgroups are defined prospectively based on expert knowledge, EEA-discovered subgroups are identified during the EEA component of the audit. In these cases, the error feature discovered should be investigated as above with a TSSA. However, unlike prospectively defined subgroups, subgroup cases identified during the EEA are even less likely to be labelled and the auditor may need to invest time and resources to carry out further targeted labelling of the audit

dataset. The risk priority score is helpful in this context, to help the auditor rationalise whether this investment is necessary. In the hip fracture detection audit, discovery of algorithmic errors in cases with abnormal bones prompted an additional labelling exercise of all hips with abnormal bones and joints to find the error rate in those cases were 50%, compared to 2.5% in the overall dataset.

Adversarial testing

While it is generally less of a concern in medical settings (where data generation and processing is heavily standardized and monitored), it can be useful to consider “worst-case” scenarios for targeted testing. The term “adversarial” here means the sort of actions that a hostile actor may take to “break” the system, but in the medical context we could consider adversarial testing roughly analogous with counterfactual reasoning - where users can explore or simulate changes in data inputs to observe how the model behaves. This can be done in a safe environment to simulate high-risk situations and their potential consequences. For example, De Grave *et al* used multiple adversarial testing approaches for a Covid-19 detection model for chest radiographs to show the model can be misled by laterality markers and shoulder positioning(35).

Unlike subgroup analyses, adversarial testing may require access to the model itself. It may also require gathering real-world examples of specific subgroups where performance is known to be poor, particularly if the subgroup is rare enough in the test data that few valid conclusions can be drawn. Alternatively, simulated data can be used. While this is more common in tabular data, recent advances in generative models can allow for the simulation of more complex data, such as images and text. With either real-world or simulated data, the purpose of adversarial testing is to better understand the prevalence and source of errors in worst-case subgroups.

Reflection

The final stage of the audit is a reflection of test results in light of the intended use and the intended impact outlined in the *scoping phase*. A final assessment of risk is formalised at this stage, risk mitigation strategies are proposed, and recommendations are made on whether the errors fall above or below threshold for continued use of the AI system. This decision will be highly specific to the clinical setting and its ability to put in place risk mitigation measures. Those

overseeing the algorithmic audit should be vigilant to deviations from the AI system's intended use. It may become apparent during testing when a mismatch between intended use and actual use has occurred, but additional auditing measures such as root cause analysis may be required to retrospectively determine whether errors were due to gaps between the intended and actual use. In any case, errors should be reported to the relevant regulatory bodies, especially if the errors found invalidates the AI system's intended use claim or the indications for use. It is also important to report errors even if adverse outcomes were mitigated through other measures in the health system (i.e., near misses), as other deployment sites may not have the same mitigation measures in place and harm which was prevented in one setting may not be prevented in another.

Risk Mitigation Measures

The feasibility and success of risk mitigation strategies will be specific to the deployment setting and requires regular review as clinical systems change over time. The measures which can be put in place also depend on who the auditor is and which aspects of the AI system and healthcare system they are able to modify. Developers may be limited by legal requirements, where any substantial changes to the AI model, deployment infrastructure, or the intended use may need to be re-evaluated by regulatory agencies.

Developer Actions

1. Retrain the AI model

The developer may decide to alter the AI system to mitigate the identified risks. Modifications could target any part of the AI system, but most commonly will involve targeted retraining of the model itself. In general, the intention would be to train an improved version of the AI model using more diverse and representative data, targeting any areas of weakness by enriching the training dataset with more examples of cases associated with errors. If further data is not available, a similar effect may be achieved by reweighting the training examples or rebalancing the training data to increase the relative value of these cases, or by producing simulated examples of these error cases.

2. Modify the model threshold

The model threshold in classification systems determines the cut-off to discriminate between positive and negative cases, and is also known as the operating point. This can be altered

without retraining the model, for example if user feedback suggests that the model produces too many false positives, then shifting the threshold can reduce these (at the expense of increasing the rate of false negatives). The operating point of a model may be pre-specified or suggested by the developers, but may also need tuning after deployment based on the specific clinical needs at a particular site.

3. Modify the instructions for use or intended use

Modifications can also affect the non-model components of the deployed infrastructure and AI workflow. This could involve changes to data acquisition and pre-processing steps, or in more extreme cases modifying the intended use of the system. Such modifications could involve excluding some types of input data from the AI system, changing how the model outputs are presented to the users, or even redefining the intended user group (for example, by increasing training requirements for users).

Clinical Actions

1. Continue use with additional human oversight

Some errors may be acceptable for continued use if the likelihood of harm is very low, or if the consequences can easily be mitigated given adequate human-oversight. Depending on the use-case, reducing the level of autonomy of the AI system and necessitating human verification may be sufficient to mitigate risks. In the FMEA, the risk prioritisation number may be informative as such errors would score low for severity and/or high for detection.
Human-oversight may be implemented for all use or reserved for certain subgroups where performance is known to be lower.

2. Modify or limit use

Where modifiable risks are identified (for example, confounding visual features such as laterality markers on chest radiographs), processes can be put in place to prevent reoccurence (in this example, by standardising placement or removal/digitization of laterality markers in chest x-ray images). Modifying the input data acquisition protocol or additional pre-processing steps integrated into the workflow to minimise the effects of spurious input data elements may be required.
If modifications are unfeasible or insufficient, limiting use of the AI system on certain input data or subgroups which are prone to errors is another option. This can be implemented if the

subgroup can be identified upstream in the clinical pathway (for example, subgroups of certain demographic, input data type, or known task-specific feature variants could be identified by the imaging technicians performing a scan) and those patients can be routed to an alternative care pathway. In the hip fracture detection audit, the risk of false positives in cases of femoral deformities was determined to require further monitoring, with a potential modification of the intended use (to exclude cases with deformities) as an appropriate mitigation action, if they were confirmed to cause a failure mode (included in thesis).
However, this option is not always feasible if the subgroups are not readily identifiable prior to analysis by the AI system.
There should also be consideration as to whether such modifications inadvertently legitimises a two-tier health system, with particular groups receiving worse care, compromising the principle of distributive justice.

3. Withdraw use

The last option is withdrawal of the AI system altogether and reverting to prior care models. Where the likelihood of errors are so severe that continued use of the algorithm is no longer safe or ethical, the only option is to stop use of the AI system until modifications can be made. The potential harms of sudden and complete withdrawal of the AI system should be weighed against the harms caused by continued use with or without modifications and limitations.

A particularly attractive alternative method to consider is the use of “hard stop” thresholds, which are common components of medical device deployments.(36) These involve pre-specified minimum performance levels, where if performance falls below the threshold during ongoing and active monitoring, then the device is immediately removed from use. These thresholds can be clearly informed by relevant organisational values (including equity and justice), and prespecification can simplify the often complex “stop use” decisions as it can allow all relevant stakeholders to be involved in these deliberations.

Algorithmic audit summary report

Findings of the medical algorithmic audit are summarised in a report which includes all collected artifacts, the FMEA, datasets, test results, risk mitigation plans and final decisions made. Any learning derived from the audit process which extends beyond the current application, should be

recorded to assist future AI evaluations/deployments. Key audit findings which carry direct implications on clinical care should also be disseminated to users. Any updates or changes made to the AI system should be made apparent to the user, ideally with reasons reported. A frequent and open dialogue of findings from the algorithmic audit summary report should be shared between developers and users.

Conclusion

AI systems for healthcare may bring significant benefits to patient care, but like any other medical intervention they also have the potential to fail and cause harm. In the case of AI algorithms, the nature of errors may be particularly difficult to discover, explain and mitigate, given the tendency for AI systems to yield unpredictable, and often subtle, errors. At a time where AI systems are being rapidly adopted into clinical practice, ensuring a framework for ongoing performance monitoring and scrutiny of error and harm is essential. These can be especially high-risk, given the deployment of AI systems often coincides with the establishment of new clinical pathways with no clear comparators for expected outcomes or standards for quality (such as the creation of new telemedical and virtual care pathways).

It is worth noting, although many AI systems are supported by evidence showing superior or equivalent performance in comparison to human experts, such monitoring of human performance is not routinely monitored in a task-specific fashion in actual clinical practice. In fact, recent clinical AI evaluations have provided valuable insights into human performance by measuring and benchmarking human performance at specific diagnostic tasks. Routine monitoring of human grader accuracy, such as those introduced by UK national screening programmes for diabetic retinopathy(37) and breast cancer screening(38), is not performed for most other clinical tasks. Gaining an understanding of human performance will not only reveal which tasks AI systems truly provide value, but will also hopefully drive the motivation to improve higher standards of care in clinicians.

The medical algorithmic audit proposed in this paper is a process to investigate and even preempt errors and harms which can be caused by AI systems. It is a general framework which promotes thoughtful interrogation of errors and unexpected results in evaluations of AI systems prior to and during real-world deployment. Performing the audit requires clinical and technical

expertise and contextual knowledge for anticipating the potential effects of the deployment environment, which may expose vulnerabilities of the algorithm and increase the likelihood of errors.

One question yet to be answered is *who* should conduct the Medical Algorithmic Audit? The skills and knowledge required to undertake such an audit crosses computational, bioinformatics and clinical skill sets, and are not currently taught in standard medical or technical curricula. In order to fulfill this responsibility, health providers need to invest in upskilling clinical personnel to oversee the piloting, deployment and ongoing monitoring of AI systems - broadly described as the science of *Algorithmovigilance*.(39) In the UK, the need to invest in digital leaders with the necessary capabilities (such as clinical information officers) has been recognised by the National Health System and Health Education England.(40,41) In Australia, the Royal Australian and New Zealand College of Radiologists have recently recommended that medical imaging departments and practices appoint a responsible radiologist with the necessary skills and knowledge to perform regular algorithmic audits of AI systems in deployment(42). In both nations, concerns have been raised that appropriately trained clinicians are rare, and that there remains significant work to be done in building an AI-ready workforce. Structured processes and guidelines such as the ones described here are necessary to accelerate the development of clinically-relevant AI quality and safety capabilities.

Ultimately, the responsibility and benefits of investigating and improving the safety of the AI system is shared between developers, healthcare decision-makers, and users and should be part of a larger oversight framework of algorithmovigilance to ensure the continued efficacy and safety of AI systems. For the medical algorithmic audit to have the highest chance of success, we advocate for the process being carried out jointly between these stakeholders, where each party enables the other in developing a deeper and more contextualised insight into the findings and possible mitigation strategies.

Acknowledgements

XL and AKD receive a proportion of their funding from the Wellcome Trust, through a Health Improvement Challenge grant (200141/Z/15/Z). BG receives funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement No 757173, project MIRA, ERC-2017-STG).

Author Contributions

All authors contributed to the conception, writing, and editing of the manuscript.

Competing Interests

BG is Scientific Advisor for Kheiron Medical Technologies, Part-time Employee and Scientific Lead of the HeartFlow-Imperial Research Team and Visiting Researcher and was a part-time Employee at Microsoft Research. The remaining authors declared no conflict of interests.

References

1. Liu X, Faes L, Kale AU, Wagner SK, Fu DJ, Bruynseels A, et al. A comparison of deep learning performance against health-care professionals in detecting diseases from medical imaging: a systematic review and meta-analysis. The Lancet Digital Health. 2019 Oct 1;1(6):e271–97.

2. Nagendran M, Chen Y, Lovejoy CA, Gordon AC, Komorowski M, Harvey H, et al. Artificial intelligence versus clinicians: systematic review of design, reporting standards, and claims of deep learning studies. BMJ. 2020 Mar 25;368:m689.

3. Wiens J, Saria S, Sendak M, Ghassemi M, Liu VX, Doshi-Velez F, et al. Do no harm: a roadmap for responsible machine learning for health care. Nat Med. 2019 Sep;25(9):1337–40.

4. Schulam P, Saria S. Can You Trust This Prediction? Auditing Pointwise Reliability After Learning. In: Chaudhuri K, Sugiyama M, editors. Proceedings of Machine Learning Research. PMLR; 2019. p. 1022–31. (Proceedings of Machine Learning Research; vol. 89).

5. Pooch EHP, Ballester P, Barros RC. Can We Trust Deep Learning Based Diagnosis? The Impact of Domain Shift in Chest Radiograph Classification. In: Thoracic Image Analysis. Springer International Publishing; 2020. p. 74–83.

6. Mahajan V, Venugopal VK, Murugavel M, Mahajan H. The Algorithmic Audit: Working with Vendors to Validate Radiology-AI Algorithms-How We Do It. Acad Radiol. 2020 Jan;27(1):132–5.

7. Liu X, The SPIRIT-AI and CONSORT-AI Working Group, Rivera SC, Moher D, Calvert MJ, Denniston AK. Reporting guidelines for clinical trial reports for interventions involving artificial intelligence: the CONSORT-AI extension [Internet]. Vol. 26, Nature Medicine. 2020. p. 1364–74. Available from: http://dx.doi.org/10.1038/s41591-020-1034-x

8. Rivera SC, The SPIRIT-AI and CONSORT-AI Working Group, Liu X, Chan A-W, Denniston AK, Calvert MJ, et al. Guidelines for clinical trial protocols for interventions involving artificial intelligence: the SPIRIT-AI extension [Internet]. Vol. 26, Nature Medicine. 2020. p. 1351–63. Available from: http://dx.doi.org/10.1038/s41591-020-1037-7

9. Winkler JK, Fink C, Toberer F, Enk A, Deinlein T, Hofmann-Wellenhof R, et al. Association Between Surgical Skin Markings in Dermoscopic Images and Diagnostic Performance of a Deep Learning Convolutional Neural Network for Melanoma Recognition. JAMA Dermatol [Internet]. 2019 Aug 14; Available from: http://dx.doi.org/10.1001/jamadermatol.2019.1735

10. Degnan AJ, Ghobadi EH, Hardy P, Krupinski E, Scali EP, Stratchko L, et al. Perceptual and Interpretive Error in Diagnostic Radiology—Causes and Potential Solutions. Acad Radiol. 2019 Jun 1;26(6):833–45.

11. Du-Harpur X, Arthurs C, Ganier C, Woolf R, Laftah Z, Lakhan M, et al. Clinically-relevant

vulnerabilities of deep machine learning systems for skin cancer diagnosis. J Invest Dermatol [Internet]. 2020 Sep 12; Available from: http://dx.doi.org/10.1016/j.jid.2020.07.034

12. Lyell D, Coiera E. Automation bias and verification complexity: a systematic review. J Am Med Inform Assoc. 2017 Mar 1;24(2):423–31.

13. Oakden-Rayner L, Dunnmon J, Carneiro G, Re C. Hidden stratification causes clinically meaningful failures in machine learning for medical imaging. In: Proceedings of the ACM Conference on Health, Inference, and Learning. New York, NY, USA: Association for Computing Machinery; 2020. p. 151–9. (CHIL '20).

14. McCradden MD, Stephenson EA, Anderson JA. Clinical research underlies ethical integration of healthcare artificial intelligence. Nat Med. 2020 Sep;26(9):1325–6.

15. Postmarket Requirements (Medical Devices and Radiation-Emitting Products) [Internet]. [cited 2020 Dec 4]. Available from: https://www.fda.gov/medical-devices/device-advice-comprehensive-regulatory-assistance/guidance-documents-medical-devices-and-radiation-emitting-products

16. Regulation (EU) 2017/745 of the European Parliament and of the Council of 5 April 2017 on medical devices, amending Directive 2001/83/EC, Regulation (EC) No 178/2002 and Regulation (EC) No 1223/2009 and repealing Council Directives 90/385/EEC and 93/42/EEC. EUR-Lex [Internet]. Official Journal of the European Union. Available from: https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A32017R0745

17. Wang P, Berzin TM, Glissen Brown JR, Bharadwaj S, Becq A, Xiao X, et al. Real-time automatic detection system increases colonoscopic polyp and adenoma detection rates: a prospective randomised controlled study. Gut. 2019 Oct;68(10):1813–9.

18. Kimmelman J, London AJ. The structure of clinical translation: efficiency, information, and ethics. Hastings Cent Rep. 2015 Mar;45(2):27–39.

19. Raji ID, Smart A, White RN, Mitchell M, Gebru T, Hutchinson B, et al. Closing the AI accountability gap: defining an end-to-end framework for internal algorithmic auditing. In: Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency. New York, NY, USA: Association for Computing Machinery; 2020. p. 33–44. (FAT* '20).

20. Shebl NA, Franklin BD, Barber N. Failure mode and effects analysis outputs: are they valid? BMC Health Serv Res. 2012 Jun 10;12:150.

21. Center for Devices, Radiological Health. PMA Labeling [Internet]. 2019 [cited 2020 Sep 29]. Available from: https://www.fda.gov/medical-devices/premarket-approval-pma/pma-labeling

22. IEC 62366-1:2015 [Internet]. 2020 [cited 2020 Sep 29]. Available from: https://www.iso.org/standard/63179.html

23. Castro DC, Walker I, Glocker B. Causality matters in medical imaging. Nat Commun. 2020 Jul 22;11(1):3673.

24. Gebru T, Morgenstern J, Vecchione B, Vaughan JW, Wallach H, Iii HD, et al. Datasheets for

Datasets [Internet]. arXiv [cs.DB]. 2018. Available from: http://arxiv.org/abs/1803.09010

25. Mitchell M, Wu S, Zaldivar A, Barnes P, Vasserman L, Hutchinson B, et al. Model Cards for Model Reporting. In: Proceedings of the Conference on Fairness, Accountability, and Transparency. New York, NY, USA: Association for Computing Machinery; 2019. p. 220–9. (FAT* '19).

26. De Fauw J, Ledsam JR, Romera-Paredes B, Nikolov S, Tomasev N, Blackwell S, et al. Clinically applicable deep learning for diagnosis and referral in retinal disease. Nat Med. 2018 Sep;24(9):1342–50.

27. Gale W, Oakden-Rayner L, Carneiro G, Bradley AP, Palmer LJ. Detecting hip fractures with radiologist-level performance using deep neural networks [Internet]. arXiv [cs.CV]. 2017. Available from: http://arxiv.org/abs/1711.06504

28. Gale W, Oakden-Rayner L, Carneiro G, Palmer LJ, Bradley AP. Producing Radiologist-Quality Reports for Interpretable Deep Learning. In: 2019 IEEE 16th International Symposium on Biomedical Imaging (ISBI 2019). 2019. p. 1275–9.

29. Sohoni NS, Dunnmon J, Angus G, Gu A, Ré C. No Subclass Left Behind: Fine-Grained Robustness in Coarse-Grained Classification Problems. In: NeurIPS [Internet]. 2020 [cited 2020 Nov 12]. Available from: https://www.semanticscholar.org/paper/8c96b865bbe1f597cf2c644e20ae46eab8e7caad

30. Ting DSW, Cheung CY-L, Lim G, Tan GSW, Quang ND, Gan A, et al. Development and Validation of a Deep Learning System for Diabetic Retinopathy and Related Eye Diseases Using Retinal Images From Multiethnic Populations With Diabetes. JAMA. 2017 Dec 12;318(22):2211–23.

31. Gulshan V, Peng L, Coram M, Stumpe MC, Wu D, Narayanaswamy A, et al. Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. JAMA - Journal of the American Medical Association. 2016;316(22):2402–10.

32. McKinney SM, Sieniek M, Godbole V, Godwin J, Antropova N, Ashrafian H, et al. International evaluation of an AI system for breast cancer screening. Nature. 2020 Jan;577(7788):89–94.

33. Ciompi F, Chung K, van Riel SJ, Setio AAA, Gerke PK, Jacobs C, et al. Towards automatic pulmonary nodule management in lung cancer screening with deep learning. Sci Rep. 7:46479.

34. Liu Y, Jain A, Eng C, Way DH, Lee K, Bui P, et al. A deep learning system for differential diagnosis of skin diseases. Nat Med. 2020 Jun;26(6):900–8.

35. DeGrave AJ, Janizek JD, Lee S-I. AI for radiographic COVID-19 detection selects shortcuts over signal [Internet]. Available from: http://dx.doi.org/10.1101/2020.09.13.20193565

36. Center for Devices, Radiological Health. Recalls, Corrections and Removals (Devices) [Internet]. [cited 2020 Nov 26]. Available from:

https://www.fda.gov/medical-devices/postmarket-requirements-devices/recalls-corrections-and-removals-devices#4

37. Diabetic eye screening: participation in the grading test and training system [Internet]. [cited 2020 Dec 9]. Available from: https://www.gov.uk/government/publications/diabetic-eye-screening-test-and-training-participation/diabetic-eye-screening-participation-in-the-grading-test-and-training-system

38. Quality Assurance Guidelines for Breast Cancer Screening Radiology - NHS Breast Screening Programme. 2011 Mar. Report No.: 2nd Editiion.

39. Embi PJ. Algorithmovigilance-Advancing Methods to Analyze and Monitor Artificial Intelligence-Driven Health Care for Effectiveness and Equity. JAMA Netw Open. 2021 Apr 1;4(4):e214622.

40. Topol E. The Topol review: preparing the healthcare workforce to deliver the digital future. Health Education England. 2019;

41. NHS. The NHS long term plan. 2019; Available from: www.longtermplan.nhs.uk

42. The Royal Australian and New Zealand College of Radiologists. Standards of Practice for Artificial Intelligence [Internet]. 2020 Jul [cited 2020 Nov 25]. Available from: https://www.ranzcr.com/whats-on/news-media/420-ranzcr-launches-world-leading-standards-for-the-use-of-ai-in-healthcare

Section 3: Explainability and ‘black box’ medical research

Deep learning models are ‘black box’ systems, meaning the mechanism of decision making is opaque to users and developers. Understanding the decisions made by AI models is broadly termed explainable AI (XAI), and the use of explainability techniques is often cited as a potential solution to various factors that are involved in the implementation gap; for example, it is common to hear that explainability can increase user trust and reduce the harm caused by unexpected or aberrant model behaviour (discussed in section 2.2) by allowing users to detect these model errors [39]. Claims like these have motivated the inclusion of XAI methods in regulatory guidelines, professional standards documents, and even in legislation [7,40,41].

However, these claims have not matched the experience of AI users. In general, AI explanations can provide useful insights into how AI models make decisions in a broad sense (sometimes called “global explainability”) but are rarely of assistance in determining the validity of individual AI decisions (termed “local explainability”). In fact, several studies have now reported that XAI can make clinical users *less likely* to recognise bad AI decisions [39,42,43], raising the possibility that explainability techniques may in fact *widen* the implementation gap.

In “**The false hope of current approaches to explainable artificial intelligence in health care”** [47] I contrast the desirable goal of XAI with the current reality; that XAI is useful for many things but can be extremely misleading for clinical users, and that clarity is required about the goals and intent of XAI policy.

The false hope of explainable AI

Publication status: In press, *Lancet Digital Health, 2021.*

Contribution: 33%

Detailed description contribution: This work was extremely collaborative, and involved all three authors (myself and Professors Beam and Ghassemi) contributing equally to the conception, planning, drafting, literature search, and editing of this work.

Certification from co-authors:

Marzyeh Ghassemi

Andrew Beam

Declaration: This publication was part of the work undertaken during and for my HDR candidature.

The false hope of current approaches to explainable artificial intelligence in health care

Marzyeh Ghassemi, Luke Oakden-Rayner, Andrew L Beam

The black-box nature of current artificial intelligence (AI) has caused some to question whether AI must be explainable to be used in high-stakes scenarios such as medicine. It has been argued that explainable AI will engender trust with the health-care workforce, provide transparency into the AI decision making process, and potentially mitigate various kinds of bias. In this Viewpoint, we argue that this argument represents a false hope for explainable AI and that current explainability methods are unlikely to achieve these goals for patient-level decision support. We provide an overview of current explainability techniques and highlight how various failure cases can cause problems for decision making for individual patients. In the absence of suitable explainability methods, we advocate for rigorous internal and external validation of AI models as a more direct means of achieving the goals often associated with explainability, and we caution against having explainability be a requirement for clinically deployed models.

Lancet Digit Health 2021; 3: e745–50

Department of Electrical Engineering and Computer Science and Institute for Medical and Evaluative Sciences, Massachusetts Institute of Technology, Cambridge, MA, USA (M Ghassemi PhD)**; Vector Institute, Toronto, ON, Canada** (M Ghassemi)**; Australian Institute for Machine Learning, University of Adelaide, Adelaide, SA, Australia** (L Oakden-Rayner)**; CAUSALab and Department of Epidemiology, Harvard T H Chan School of Public Health, Boston, MA, USA** (A L Beam PhD)**; Department of Biomedical Informatics, Harvard Medical School, Boston, MA, USA** (A L Beam)

Correspondence to:
Dr Andrew L Beam, Department of Epidemiology, Harvard T H Chan School of Public Health, Boston, MA 02115, USA
andrew_beam@hms.harvard.edu

Introduction

Artificial intelligence (AI), powered by advances in machine learning, has made substantial progress across many areas of medicine in the past decade.[1–5] Given the increasing ubiquity of AI techniques, a new challenge for medical AI is its so-called black-box nature, with decisions that seem opaque and inscrutable. In response to the uneasiness of working with black boxes, there is a growing chorus of clinicians, lawmakers, and researchers calling for explainable AI models for high-risk areas such as health care.[6,7]

Although precise technical definitions of explainability lack consensus,[8,9] many high-level, less precise definitions have been put forth by various stakeholders. For example, the General Data Protection Regulation laws in the EU state that all people have the right to "meaningful information about the logic behind automated decisions using their data".[10,11] Similar discussions have taken place in the clinical literature, in which it has been argued that clinicians might feel uncomfortable with black-box AI,[12] leading to recommendations[13] that AI should be explainable in a way that clinical users can understand. Indeed, Tonekaboni and colleagues report that surveyed clinicians "viewed explainability as a means of justifying their clinical decision-making".[14]

We believe that the desire to engender trust through current explainability approaches represents a false hope: that individual users or those affected by AI will be able to judge the quality of an AI decision by reviewing a local explanation (that is, an explanation specific to that individual decision[8]). These stakeholders might have misunderstood the capabilities of contemporary explainability techniques—they can produce broad descriptions of how the AI system works in a general sense but, for individual decisions, the explanations are unreliable or, in some instances, only offer superficial levels of explanation. In practice, explanations can be extremely useful when applied to global AI processes, such as model development, knowledge discovery, and audit, but they are rarely informative with respect to individual decisions.

As such, we suggest that end users of explainable AI, including clinicians, lawmakers, and regulators, be aware of the limitations of explainable AI as it currently exists, especially as it relates to policy, use, and reporting. We argue that if the desire is to ensure that AI systems can operate safely and reliably, the focus should be on rigorous and thorough validation procedures.

Current approaches to explainable AI

Attempts to produce human-comprehensible explanations for machine learning decisions have typically been divided into two categories: inherent explainability and post-hoc explainability.

For machine learning models for which the input data are of limited complexity and clearly understandable, quantifying the relationships between these simple inputs and the outputs of the model is termed inherent explainability. An example of this would be in a linear regression model, where a coefficient measures the strength and direction of the relationship between the weight of a car and the fuel efficiency. The coefficient itself characterises the decision in an understandable way by describing how much each additional kilogram reduces fuel efficiency on average.

The intuitive simplicity of inherently explainable models is appealing, but even these explanations are hampered by the presence of unrecognised confounders. Work in the human–computer interaction community has identified that increased transparency can hamper users' ability to detect sizable model errors and correct for them, "seemingly due to information overload",[15] even for clear-box or inherently explainable models. Further work has found that even data scientists "over-trust and misuse interpretability tools" and that few such experts were able to accurately describe visualisations output by interpretability tools.[16]

In contrast to inherently explainable models, in many modern AI use cases, the data and models are too complex and high-dimensional to be easily understood; they cannot be explained by a simple relationship

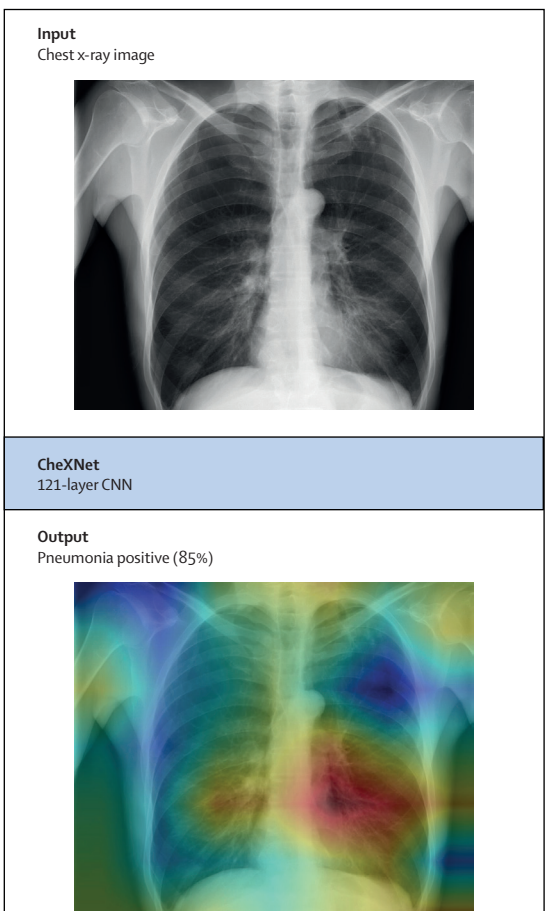


***Figure 1*: Heat map produced by a post-hoc explanation method for a deep learning model designed to detect pneumonia in chest x-rays**
Brighter colours (red) indicate regions with higher levels of importance according to the deep neural network, and darker colours (blue) indicate regions with lower levels of importance. Reproduced with permission from Rajpurkar et al.[21] CNN=convolutional neural network.

See Online for appendix

between inputs and outputs. Examples include models designed to analyse images, text, and sound data. In these scenarios, the focus has been on attempting to dissect the model's decision making procedure, a process called post-hoc explainability. To show post-hoc explainability, we use medical imaging as an illustrative example and explore the most commonly used form of post-hoc explainability in this setting: heat maps. Heat maps (or saliency maps)[17–19] highlight how much each region of the image contributed to a given decision and are illustrative because they provide a simple means of understanding some of the limitations of post-hoc explainability techniques. Although they are popular for medical imaging models, they are well known to be problematic in the broader explainability literature.[20]

As an example, the saliency map shown in figure 1, from Rajpurkar and colleagues,[21] highlights the areas of the image deemed most important for the diagnosis of pneumonia. Even the hottest parts of the map contain both useful and non-useful information (from the perspective of a human expert), and simply localising the region does not reveal exactly what it was in that area that the model considered useful. The clinician cannot know if the model appropriately established that the presence of an airspace opacity was important in the decision, if the shapes of the heart border or left pulmonary artery were the deciding factor, or if the model had relied on an inhuman feature, such as a particular pixel value or texture that might have more to do with the image acquisition process than the underlying disease.

This interpretability gap of explainability methods relies on humans to decide what a given explanation might mean. Unfortunately, the human tendency is to ascribe a positive interpretation: we assume that the feature we would find important is the one that was used (this is an example of a famously harmful cognitive error called confirmation bias). This problem is well summarised by computer scientist Cynthia Rudin: "You could have many explanations for what a complex model is doing. Do you just pick the one you 'want' to be correct?".[22] The ability of localisation methods to mislead human users is compellingly demonstrated by Adebayo and colleagues,[20] who show that even untrained networks can produce saliency maps that appear reassuring (appendix). Moreover, Gu and Tresp[23] showed that common visual explanations remain unchanged even when precise modifications are made to the input that substantially alter the model's predictions (a process known as an adversarial attack), even when those attacks lead to incorrect model predictions (figure 2). It is hard to credit the explanatory ability of a technique that appears believable even when the model is wrong or even completely untrained.

The interpretability gap exists beyond imaging as well. As an example, we see similar problems with contextual language models such as SciBERT,[24] trained on seemingly innocuous sources such as PubMed, which have been shown to have deeply problematic associations about gender and race.[25] Although explanations for language tend to revolve around highlighting the words in the text that contributed to the decision, this does not reveal the associative meaning the model has learned for those words. As with heat maps, the human tendency is to assume that a model has used words in the same way we would. However, deeper investigation often reveals that these models rely on unacceptable shortcuts, such as strongly associating the word doctor with maleness and using this reductionist interpretation to inform decision making.

Beyond heat maps, many other approaches have been developed to produce explanations in complex medical data, including methods such as feature visualisation and prototypical comparisons. Feature visualisation involves the production of synthetic inputs that most strongly activate specific parts of a machine learning

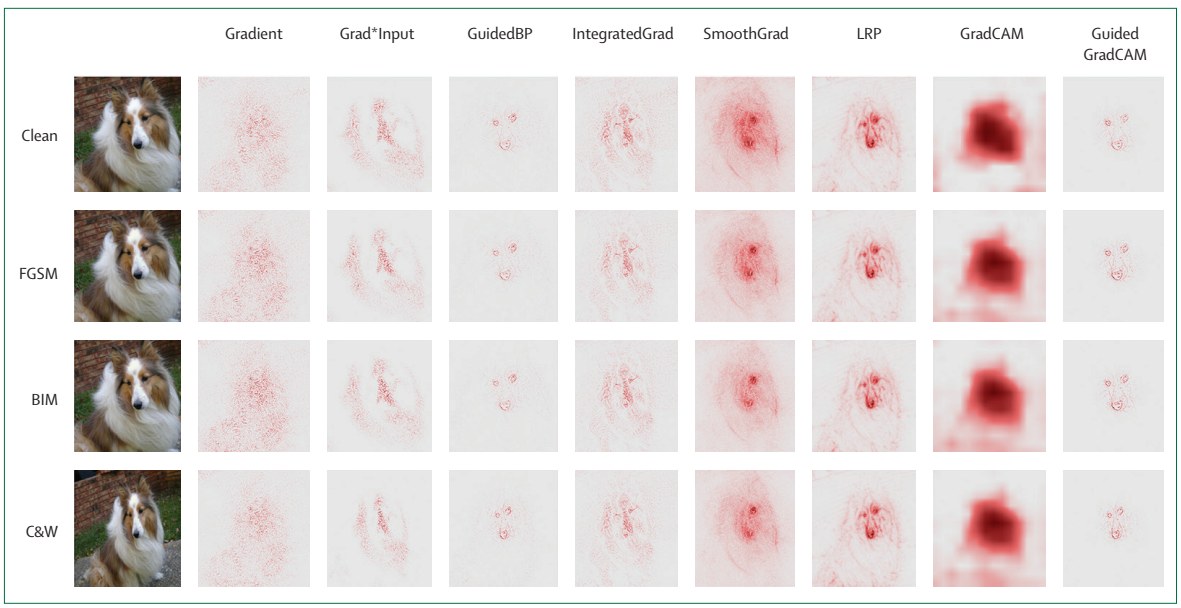


***Figure 2*: Saliency maps produced by popular methods**
Each column shows a different type of explainability method that highlights the most relevant pixels in the images below, which, in each row, are subject to different adverserial pertubations. The top row shows the correctly classified image and saliency maps, and rows 2 to 4 show incorrectly classified images after adversarial perturbations. Reproduced with permission from Gu and Tresp.[23] BIM=basic iteractive method. C&W=Carlini & Wagner. FGSM=fast gradient sign method. GradCAM=gradient-weighted class activation mapping. GuidedBP=guided backpropagation. LRP=layerwise relevance propagation.

model.[26] Each model decision can then be described as a combination of a series of features that were detected in the input. In practice, these synthetic inputs rarely correspond exactly to specific human-interpretable features and are subject to the exact same concerns as heat maps: if a synthetic input looks roughly like a feature a human would use to make a decision (for example, a fur-like texture feature in a dog-detecting AI model), a human must still interpret whether this implies the model made a good decision.

These concerns also extend to other well known post-hoc explanation methods such as locally interpretable model-agnostic explanations (LIME)[27] and Shapley values (SHAP).[28] LIME seeks to understand decisions at the individual level by permuting the input example (altering it in minor ways) and identifying which alterations were most likely to change the decision. In the case of image analysis, this is done by occluding parts of the image, the explanation consisting of a heat map that indicates the image components that were most important for the decision. Such explanations suffer from interpretability gaps in the same way as saliency mapping. Methods such as LIME and SHAP are generic and not specific to images and are routinely used on a wide variety of health-care data, including structured data from electronic health-care records[29] and electroencephalogram waveform data.[30]

Prototypical explanations are interesting in that they are generally considered to be a form of inherent explainability. The model is not only trained for the task itself, but to also identify prototypical elements of each class and then quantify how much of each component it identified for the given decision. Examples include comparing the relevant parts of an image (such as the beak and claws of a bird) to a prototype,[31] producing a text-based description of the decision by referencing canonical descriptive features,[32] or identifying a training instance that is most similar to a test instance according to the trained model.[33] This type of learned explanation has only been recently proposed and has yet to be applied broadly, but still requires human interpretation (ie, were the right canonical elements selected? Was the proportion of each element appropriate?).

All of these examples reveal another major challenge: explanations have no performance guarantees. Indeed, the performance of explanations is rarely tested at all, and most tests that are done rely on heuristic measures rather than explicitly scoring the explanation from a human perspective.[34] This is problematic because explanations, such as those shown in figures 1 and 2, are only approximations to the model's decision procedure and therefore do not fully capture how the underlying model will behave. As such, using post-hoc explanations to assess the quality of model decisions adds an additional source of error—not only can the model be right or wrong, but so can the explanation. Rudin takes this further, saying that post-hoc explanations "must be wrong"; that they are by definition not completely faithful to the original model and must be less accurate with respect to the primary task.[35] In this context, should researchers prefer the full, complex model, which, as humans, we cannot understand but has a high, validated

performance or do we seek to modify that performance with an explanation mechanism, potentially resulting in diminished and unvalidated accuracy?

What are explanations for?

These limitations do not render explainability methods useless, but they do challenge the use of these techniques for certain purposes. If we look at the policy positions and user preferences mentioned earlier, or the intuitive expectations that AI is made explainable, we see a desire to generate trust and inform the choices of individual users or the subjects of AI decision making. However, on an individual level, the explanations we can produce for the behaviour of complex AI systems are often confusing or even misleading. Selbst and Barocas[36] state that, although explainability methods can provide some insight into the decision making process of models, they rarely elucidate whether a given decision was sensible or not. Selbst and Barocas distinguish between explainability techniques that provide descriptive accounts of how the model behaved and normative evaluations that can answer whether that behaviour was justified. Although most discussions and policies call for normative evaluations, current techniques are only capable of descriptive accounts and it is our own intuition that often "serves as the unacknowledged bridge" between the two.[36]

In the example of heat maps, the important question for users trying to understand an individual decision is not where the model was looking but instead whether it was reasonable that the model was looking in this region. By conflating these questions and allowing intuition to bridge the gap, there is a serious risk of introducing harmful biases into decision making. There is a great deal of evidence that humans tend to over-trust computer systems,[37–39] and evidence suggests models that use explainability techniques can hamper people's ability to detect when a model makes serious mistakes[15] or unreasonably increase their confidence in an algorithmic decision,[40,41] giving the veneer of authenticity and resulting in decreased vigilance and auditing of such systems.

This tendency is particularly problematic as another goal of explainability is to detect and avoid algorithms biased towards certain populations.[25] Many systematic biases that reflect societal prejudices (eg, discriminatory policies against women and minority ethnic groups) are encoded in the data from which the AI system learns. Left unchecked, an AI system could operationalise these biases on a large scale. It is implied that explainability could allow us to catch discriminatory behaviour more readily. Unfortunately, as outlined above, this possibility is not reflected in the current state of explainability research, and reliance on explanations might even decrease our vigilance for these behaviours.

Rather than seeing explainability techniques as producing valid, local explanations to justify the use of model predictions, it is more realistic to view these methods as global descriptions of how a model functions. If, for example, a clinical diagnostic model appears to perform well in a specific test set but the heat maps show that the model is consistently distracted by regions of the images that cannot logically inform the diagnosis, then this finding can indicate that the test set itself is flawed and that further forensic investigation is required. An example of this use was when explanatory heat maps revealed that an AI model trained to detect skin cancer was focusing more on the surgical skin markings present on the images rather than the skin lesions.[42] Similarly, there have been notable successes in using explainability methods to aid in the discovery of knowledge, for example, when heat maps were used to identify novel features of diabetic retinopathy progression in ophthalmological fundal eye examination[43] and new radiographic features that are predictive of knee pain.[44] In this sense, we can see it is the aggregate behaviour of these explanations that is informative, not the unquantifiable effect a single reassuring or aberrant explanation will have on an individual prediction.

Better and more equitable outcomes

Although explanations cannot provide a normative evaluation of our models, that does not mean we are forced to accept their black-box predictions without scrutiny. As we have argued, it is the aggregated behaviour of the models that can be informative and, as such, the only effective way to justify the decisions of AI systems is thorough, careful, meticulous safety and validation efforts. Instead of requiring local explanations from a complicated AI system, we should advocate for thorough and rigorous validation of these systems across as many diverse and distinct populations as possible, showing that patient and health-care outcomes are improved and that marginalised groups are not disproportionately affected by any given system.

The medical system is already extremely adept at evaluation and validating various kinds of black-box systems, as many drugs and devices function, in effect, as black boxes. An often cited example is acetaminophen, which, despite having been used for more than a century, has a mechanism of action that remains only partially understood.[45] Despite competing explanations for how acetaminophen works, we know that it is a safe and effective pain medication because it has been extensively validated in numerous randomised controlled trials (RCTs). RCTs have historically been the gold-standard way to evaluate medical interventions, and it should be no different for AI systems. In recognition of this, many RCT reporting guidelines are being updated to incorporate AI-specific recommendations.[46]

RCTs are not the only mechanism used in health technology assessment to ensure safety, efficacy, and equity. As an example, for an investigation into racial bias in a machine learning system,[47] even completely transparent understanding of the algorithm in question did not reveal the racial bias inherent to the model because it was the problem formulation itself that was

flawed. Instead, it was an aggregate analysis of the inputs, outputs, and outcomes associated with the model that identified the bias. In this context, explainability techniques can serve as a valuable tool for analysis and an adjunct to algorithmic audit,[48] for which the appropriate audience for explanations is not the users or subjects of AI, but rather the developers, auditors, and regulators of these systems.

Conclusions

AI will have an extraordinary impact on medicine in the coming decades, and we should do all we can to ensure that this technology is implemented in a way that maximises patient benefit. However, despite its intuitive appeal, explainability for patient-level decision making is unlikely to advance these goals in meaningful ways. Explainability methods cannot yet provide reassurance that an individual decision is correct, increase trust among users, nor justify the acceptance of AI recommendations in clinical practice.

That is not to say that explainability methods have no role in AI safety. These methods are incredibly useful for model troubleshooting and systems audit, both of which can be used to improve model performance or identify common failure modes or biases. Current explainability methods should be seen as tools for developers and auditors to interrogate their models and, unless there are substantial advances in explainable AI, we must treat these systems as black boxes, justified in their use not by just-so rationalisations, but instead by their reliable and experimentally confirmed performance.

Presently, the hope for human-comprehensible explanations for complex, black-box machine learning algorithms that can be used safely for bedside decision making remains an open challenge. In light of this challenge, we strongly recommend that health-care workers exercise appropriate caution when using explanations from an AI system and urge regulators to be judicious in listing explanations among the requirements needed for clinical deployment of AI.

Contributors

All authors contributed equally to the conception, writing, and editing of the Viewpoint, and had final responsibility for the decision to submit for publication.

Declaration of interests

We declare no competing interests.

Acknowledgments

The authors wish to thank Pranav Rajpurkar for the permission to reprint figure 1, Jindong Gu for permission to reprint figure 2, and Julius Adebayo for permission to use a modified figure in the appendix. ALB was supported by the National Heart, Lung, and Blood Institute (7K01HL141771-02).

References

1 Topol EJ. High-performance medicine: the convergence of human and artificial intelligence. *Nat Med* 2019; **25:** 44–56.

2 Ghassemi M, Naumann T, Schulam P, Beam AL, Chen IY, Ranganath R. A review of challenges and opportunities in machine learning for health. *arXiv* 2019; published online Dec 5. https://arxiv.org/abs/1806.00388 (preprint).

3 Yu K-H, Beam AL, Kohane IS. Artificial intelligence in healthcare. *Nat Biomed Eng* 2018; **2:** 719–31.

4 Beam AL, Kohane IS. Big data and machine learning in health care. *JAMA* 2018; **319:** 1317–18.

5 Beam AL, Kohane IS. Translating artificial intelligence into clinical care. *JAMA* 2016; **316:** 2368–69.

6 Gastounioti A, Kontos D. Is it time to get rid of black boxes and cultivate trust in AI? *Radiol Artif Intell* 2020; **2:** e200088.

7 Reyes M, Meier R, Pereira S, et al. On the interpretability of artificial intelligence in radiology: challenges and opportunities. *Radiol Artif Intell* 2020; **2:** e190043.

8 Doshi-Velez F, Kim B. Towards a rigorous science of interpretable machine learning. *arXiv* 2017; published online Feb 28. http://arxiv.org/abs/1702.08608 (preprint).

9 Lipton ZC. The mythos of model interpretability. *Commun ACM* 2018; **61:** 36–43.

10 European Parliament. Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46. *OJEU* 2016; **59:** 294.

11 Miller K. AI decisions: do we deserve an explanation? June 29, 2020. https://www.futurity.org/ai-decisions-right-to-explanation-2394872-2/ (accessed Sept 9, 2021).

12 Wang F, Kaushal R, Khullar D. Should health care demand interpretable artificial intelligence or accept "black box" medicine? *Ann Intern Med* 2020; **172:** 59–60.

13 Cutillo CM, Sharma KR, Foschini L, Kundu S, Mackintosh M, Mandl KD. Machine intelligence in healthcare-perspectives on trustworthiness, explainability, usability, and transparency. *NPJ Digit Med* 2020; **3:** 47.

14 Tonekaboni S, Joshi S, McCradden MD, Goldenberg A. What clinicians want: contextualizing explainable machine learning for clinical end use. *arXiv* 2019; published online May 13. http://arxiv.org/abs/1905.05134 (preprint).

15 Poursabzi-Sangdeh F, Goldstein DG, Hofman JM, Vaughan JW, Wallach H. Manipulating and measuring model interpretability. *arXiv* 2021; published online Aug 15. https://arxiv.org/abs/1802.07810 (preprint).

16 Kaur H, Nori H, Jenkins S, Caruana R, Wallach H, Vaughan JW. Interpreting interpretability: understanding data scientists' use of interpretability tools for machine learning. In: Proceedings of the 2020 CHI conference on human factors in computing systems. New York, NY, USA: Association for Computing Machinery, 2020: 1–14.

17 Selvaraju RR, Cogswell M, Das A, Vedantam R, Parikh D, Batra D. Grad-cam: visual explanations from deep networks via gradient-based localization. In: Proceedings of the IEEE international conference on computer vision. Cambridge, MA, USA: Institute of Electrical and Electronics Engineers, 2017: 618–26.

18 Tulio Ribeiro M, Singh S, Guestrin C. "Why should I trust you?": explaining the predictions of any classifier. *arXiv* 2016; published Aug 9. https://arxiv.org/abs/1602.04938 (preprint).

19 Lundberg SM, Lee S-I. A unified approach to interpreting model predictions. *Adv Neural Inf Process Syst* 2017; **30:** 4765–74.

20 Adebayo J, Gilmer J, Muelly M, Goodfellow I, Hardt M, Kim B. Sanity checks for saliency maps. *Adv Neural Inf Process Syst* 2018; **31:** 9505–15.

21 Rajpurkar P, Irvin J, Zhu K, et al. CheXNet: radiologist-level pneumonia detection on chest X-rays with deep learning. *arXiv* 2017; published online Nov 14. http://arxiv.org/abs/1711.05225 (preprint).

22 Bornstein AM. Is artificial intelligence permanently inscrutable? Sept 1, 2016. http://nautil.us/issue/40/learning/is-artificial-intelligence-permanently-inscrutable (accessed Feb 27, 2020).

23 Gu J, Tresp V. Saliency methods for explaining adversarial attacks. *arXiv* 2019; published online Aug 22. http://arxiv.org/abs/1908.08413 (preprint).

24 Beltagy I, Lo K, Cohan A. SciBERT: a pretrained language model for scientific text. *arXiv* 2019; published online Sept 10. https://arxiv.org/abs/1903.10676 (preprint).

25 Zhang H, Lu AX, Abdalla M, McDermott M, Ghassemi M. Hurtful words: quantifying biases in clinical contextual word embeddings. In: Proceedings of the ACM conference on health, inference, and learning. New York, NY, USA: Association for Computing Machinery, 2020: 110–20.

26 Olah C, Satyanarayan A, Johnson I, et al. The building blocks of interpretability. *Distill* 2018; **3:** e10.

27 Biecek P, Burzykowski T. Local interpretable model-agnostic explanations (LIME). In: Explanatory model analysis. New York, NY, USA: Chapman and Hall/CRC, 2021: 107–23.

28 Slack D, Hilgard S, Jia E, Singh S, Lakkaraju H. Fooling LIME and SHAP: adversarial attacks on post hoc explanation methods. In: Proceedings of the AAAI/ACM conference on AI, ethics, and society. New York, NY, USA: Association for Computing Machinery, 2020: 180–86.

29 Khedkar S, Gandhi P, Shinde G, Subramanian V. Deep learning and explainable AI in healthcare using EHR. In: Dash S, Acharya BR, Mittal M, Abraham A, Kelemen A, eds. Deep learning techniques for biomedical and health informatics. Cham, Germany: Springer International Publishing, 2020: 129–48.

30 Alsuradi H, Park W, Eid M. Explainable classification of EEG data for an active touch task using Shapley values. In: HCI international 2020—late breaking papers: multimodality and intelligence. Cham, Germany: Springer International Publishing, 2020: 406–16.

31 Chen C, Li O, Tao D, Barnett A, Rudin C, Su JK. This looks like that: deep learning for interpretable image recognition. *Adv Neural Inf Process Syst* 2019; **32:** 8930–41.

32 Gale W, Oakden-Rayner L, Carneiro G, Palmer LJ, Bradley AP. Producing radiologist-quality reports for interpretable deep learning. *arXiv* 2018; published online June 1. https://arxiv.org/abs/1806.00340 (preprint).

33 Schmaltz A, Beam A. Exemplar auditing for multi-label biomedical text classification. *arXiv* 2020; published online April 7. http://arxiv.org/abs/2004.03093 (preprint).

34 Gilpin LH, Bau D, Yuan BZ, Bajwa A, Specter M, Kagal L. Explaining explanations: an overview of interpretability of machine learning. In: IEEE 5th International Conference on Data Science and Advanced Analytics (DSAA). Cambridge, MA, USA: Institute for Electrical and Electronics Engineers, 2018: 80–89.

35 Rudin C. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nat Mach Intell* 2019; **1:** 206–15.

36 Selbst AD, Barocas S. The intuitive appeal of explainable machines. *Fordham Law Rev* 2018; **87:** 1085–139.

37 Skitka LJ, Mosier KL, Burdick M. Does automation bias decision-making? *Int J Hum Comput Stud* 1999; **51:** 991–1006.

38 Lyell D, Coiera E. Automation bias and verification complexity: a systematic review. *J Am Med Inform Assoc* 2017; **24:** 423–31.

39 Howard A. Are we trusting AI too much? In: Proceedings of the 2020 ACM/IEEE International Conference on Human-Robot Interaction. New York, NY, USA: Association for Computing Machinery, 2020: 1.

40 Ghassemi M, Pushkarna M, Wexler J, Johnson J, Varghese P. ClinicalVis: supporting clinical task-focused design evaluation. *arXiv* 2018; published online Oct 13. http://arxiv.org/abs/1810.05798 (preprint).

41 Eiband M, Buschek D, Kremer A, Hussmann H. The impact of placebic explanations on trust in intelligent systems. In: Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems. New York, NY, USA: Association for Computing Machinery, 2019: 1–6.

42 Winkler JK, Fink C, Toberer F, et al. Association between surgical skin markings in dermoscopic images and diagnostic performance of a deep learning convolutional neural network for melanoma recognition. *JAMA Dermatol* 2019; **155:** 1135–41.

43 Arcadu F, Benmansour F, Maunz A, Willis J, Haskova Z, Prunotto M. Deep learning algorithm predicts diabetic retinopathy progression in individual patients. *NPJ Digit Med* 2019; **2:** 92.

44 Pierson E, Cutler DM, Leskovec J, Mullainathan S, Obermeyer Z. An algorithmic approach to reducing unexplained pain disparities in underserved populations. *Nat Med* 2021; **27:** 136–40.

45 Kirkpatrick P. New clues in the acetaminophen mystery. *Nat Rev Drug Discov* 2005; **11:** 883.

46 Liu X, Faes L, Calvert MJ, Denniston AK. Extension of the CONSORT and SPIRIT statements. *Lancet* 2019; **394:** 1225.

47 Obermeyer Z, Powers B, Vogeli C, Mullainathan S. Dissecting racial bias in an algorithm used to manage the health of populations. *Science* 2019; **366:** 447–53.

48 Raji ID, Smart A, White RN, et al. Closing the AI accountability gap: defining an end-to-end framework for internal algorithmic auditing. *arXiv* 2020; published online Jan 3. http://arxiv.org/abs/2001.00973 (preprint).

Section 4: A thorough preclinical evaluation of an AI model

In sections 1 to 3, I cover a range of issues which can lead to poor performance in AI models when they are deployed in practice, and offer several solutions to improve preclinical testing and close the implementation gap.

In "**Validation and algorithmic audit of a deep learning system for the detection of proximal femoral fractures in emergency department patients"** (in press), I bring together these topics to demonstrate an example of a rigorous pre-clinical evaluation of AI. While I do not believe that this approach can replace true pre-deployment clinical trials, I show how the use of high quality, well characterised data, the application of appropriate methods to estimate human baseline performance, and algorithmic auditing can reveal implementation gaps that would otherwise go unrecognised.

Validation and algorithmic audit of a deep learning system for the detection of proximal femoral fractures in emergency department patients

Publication status: In press, *Lancet Digital Health, 2021.*

Contribution: 50%

Detailed description contribution: I was responsible for planning and study design, data collection for the local data, data cleaning for all of the data, recruiting for and performing the reader study, performing the analysis, and drafting/editing the manuscript.

The deep learning model was developed for previous work by co-author William Gale, who was also responsible for running the model on the validation datasets and producing the saliency maps. Co-authors Drs Thomas Bonham and Matthew Lungren were responsible for the collection of the external validation dataset at Stanford University Hospital. All remaining authors were responsible for study design, as well as editing the manuscript.

Certification from co-authors:

Professor Lyle Palmer | Andrew Bradley | Gustavo Carneiro

Matthew Lungren | William Gale | Thomas Adam Bonham

Declaration: This publication was part of the work undertaken during and for my HDR candidature.

Validation and algorithmic audit of a deep learning system for the detection of proximal femoral fractures in emergency department patients

Luke Oakden-Rayner*[1,2], William Gale[2,3], Thomas A. Bonham[4], Matthew P. Lungren[4,5], Gustavo Carneiro[2], Andrew P. Bradley[6], Lyle J. Palmer[1,2]

*corresponding author

Degrees:

L Oakden-Rayner: MBBS, W Gale: BSc (Hons), TA Bonham: BS, MP Lungren: MD (associate professor), G Carneiro: PhD (Full professor), AP Bradley: PhD (Full professor), LJ Palmer: PhD (Full Professor)

Affiliations:

1. School of Public Health, University of Adelaide, Adelaide, Australia
2. Australian Institute for Machine Learning, University of Adelaide, Adelaide, Australia
3. School of Computer Science, University of Adelaide, Adelaide, Australia
4. Stanford University School of Medicine Department of Radiology
5. Stanford Artificial Intelligence in Medicine and Imaging Center
6. Science and Engineering Faculty, Queensland University of Technology, Brisbane, Australia

Abstract

Background: Proximal femoral fractures are a serious clinical and public health issue associated with substantial morbidity and early mortality. Artificial intelligence may offer improved

diagnostic accuracy, but typical approaches to testing can underestimate the risks of these systems.

Methods: We present a thorough preclinical evaluation of an artificial intelligence system intended to detect proximal femoral fractures in frontal x-ray films in emergency department patients, including a reader study, an external validation, and an algorithmic audit.

Findings: The artificial intelligence system demonstrates very good summary performance (ROC AUC = 0.994, 95% CI = 0.988 to 0.999 vs radiologist SROC AUC = 0.969, 95% CI = 0.960 to 0.978), but a thorough evaluation identifies several barriers to safe deployment including a significant shift in the model operating point on external validation, and an increased error rate on cases with abnormal bones, such as those with Paget's disease and femoral head deformities.

Interpretation: Thorough pre-clinical evaluation of artificial intelligence models, in particular including algorithmic auditing, can reveal unexpected and potentially harmful behaviour even in high performance artificial intelligence systems, which can inform future clinical testing and deployment decisions.

Funding: Nil.

Research in context

Evidence before this study: We searched Google Scholar on 10th December 2019, for literature published at study inception, with no language restrictions, on: (a) deep learning-based detection of hip fractures using the keywords “hip fracture” or “proximal femoral fracture”, and “deep learning” or “artificial intelligence”, and (b) algorithmic audits of deep learning studies using the keywords “deep learning” or “artificial intelligence” and “audit”. The literature on hip fracture detection using deep learning models was limited. The majority of studies reported internal performance of the AI model only. There was a single reader study identified, which estimated human performance from a single clinician and performed no external validation. There were no studies reporting further analysis into unexpected model behaviour or failure modes. Further, no audits of medical AI systems have been reported.

Added value of this study: This study demonstrates a thorough preclinical evaluation of a high performance medical artificial intelligence system (trained to detect proximal femoral fractures on plain film imaging). Despite extremely high performance, outperforming human experts in the task of proximal femoral fracture detection, more thorough evaluation including algorithmic auditing demonstrated unexpected and potentially harmful algorithmic behaviour.

Implications of all the available evidence: Thorough evaluation of AI systems, including algorithmic auditing, can identify barriers to safe AI deployment which may not be appreciated during standard preclinical testing, and which could cause significant harm if left unrecognised.

Regulators, medical governance bodies, and professional groups should consider the need for more comprehensive preclinical testing of AI prior to clinical deployment.

Introduction

Hip fractures, or more specifically proximal femoral fractures, present a significant global clinical and public health challenge. In the elderly, proximal femoral fractures are the second most frequent cause of hospitalisation and are among the most common causes of morbidity and long-term mortality(1), with a lifetime risk of 17.5% for women and 6% for men(2). Up to 10% of patients with suspected proximal femoral fractures are not diagnosed on the initial pelvic x-ray study and undergo further diagnostic imaging(3), which may include additional x-rays, nuclear medicine bone scans, computed tomography (CT), and/or magnetic resonance imaging (MRI). Of those patients undergoing additional imaging, only around a third ultimately demonstrate a fracture(3,4). Not only does this further imaging increase the diagnostic costs, burden on doctors and patients, and resource utilisation, but these “occult fractures” may also lead to delayed or missed diagnoses and concomitant worse patient outcomes, including increased mortality rate(5,6), length of hospitalisation(7), and cost of care(8).

Improved diagnostic accuracy using x-rays taken at first clinical presentation could plausibly reduce both costs and harms. Many studies have reported that artificial intelligence (AI) systems may exceed human performance for certain diagnostic tasks(9). In order to reduce the rates of misdiagnosis or incomplete diagnosis of the initial radiograph in an emergency department we

have developed a deep learning-based proximal femoral fracture detection model with exceptional performance characteristics(10). In the current study, we evaluate the performance of the deep learning model, and compare this against the current standard of care (clinical radiologists) in a multireader multicase (MRMC) study.

The performance of deep learning models for medical image analysis has been reported in many preclinical studies(11), yet almost no clinical trials have been performed to show that these results translate into clinical practice (9). Historically, computer aided diagnosis systems have performed unexpectedly poorly in the clinical setting, despite promising preclinical evaluations(12), which has been called the “implementation gap”(13). A number of factors are expected to be responsible for this poor clinical performance, including the misapplication of models outside of intended use-cases(14,15), a variable ability to generalise to new clinical environments(16–19), statistical flaws when estimating the pooled performance and variability of human readers(20), and the presence of unidentified poor performance in clinically important subsets of cases(21). All of these factors, excepting the role of model misapplication, can be evaluated to some extent prior to clinical testing. Despite this, the majority of preclinical AI research to date has not addressed these concerns. External validation, a mechanism to assess the ability of a model to generalise to new environments, was only performed in around one third of studies considered in a recent systematic review(11). Recently, formal algorithmic auditing has been proposed(22) as a mechanism to identify and mitigate any sources of unexpected machine learning model behaviour.

We perform a thorough preclinical evaluation of a previously developed high-performance proximal femoral fracture model(10), intended to reflect current “best practice” for preclinical (i.e., prior to clinical trial) assessment, including several key components:

1) A multi-reader multi-case study design which is adequately powered to determine the relative performance of the AI model and the humans experts,
2) an external validation of the model on international data to attempt to replicate the results and identify any challenges for generalisation to new clinical sites,
3) an algorithmic audit to identify any unexpected behaviour of the deep learning model and to estimate the likelihood of a gap between pre-clinical performance and the safety of a clinical deployment.

Methods

Deep learning model

The deep learning model evaluated in this study was developed previously and has been described in detail (10). Briefly, the model consists of a DenseNet architecture (23) with 172 layers, trained on a development dataset which had no patient overlap with the study datasets, consisting of 45,786 unilateral hip images with a fracture prevalence of 11%.

Primary validation dataset

A large local dataset was obtained from the Royal Adelaide Hospital (RAH), a tertiary teaching hospital in South Australia which services adult patients only (age > 16 years). The RAH dataset included all frontal pelvis x-rays ordered between 2005 and 2015 as part of standard clinical care, obtained using a wide variety of x-ray equipment. Visual exclusion criteria included studies with no frontal pelvis film, as well as cases with prior surgical intervention with implanted metalwork. Hips containing metalwork were excluded on a per-hip basis (i.e., if only the left hip contained metal, the right hip was still included). Fractures in post-operative hips were thought to represent a visually distinct class of injury which would require an intentional training approach and specific dataset to detect with good performance, thus it was considered more clinically useful to train a model that could detect fractures in pre-operative hips. This model is not intended to function in hips with metalwork in situ.

These visual assessments to identify cases to exclude were performed by a series cascade of "helper" AI models developed and validated during earlier work, with human review of all included films to ensure appropriateness(10).

The primary validation (PV) dataset was randomly selected (at the patient level) from the emergency department cases in the RAH dataset. A total of 4,577 unilateral hip x-rays were selected, including 640 proximal femoral (hip) fractures. The ground truth for hip fracture status was determined through a combination of x-ray reports, follow-up imaging (with CT or MRI scans), and surgical records, with a follow-up period of a minimum of 6 months. Mortality records were also searched, but revealed no further cases of proximal femoral hip fractures. The

majority of proximal femoral hip fracture cases were surgically validated (91.5%), meaning the patients were surgically treated for fracture. The remainder of the patients either did not receive surgery (i.e., they died prior to surgery or were palliated), or they were transferred to other institutions prior to treatment.

The PV dataset was intended to investigate the application of our model to unseen clinical cases and hence these cases were not available to the model during training (the remainder of the RAH set was used for model development, called the Dev dataset). Emergency department referrals were chosen for inclusion in the PV dataset as this was considered the most clinically challenging setting, i.e., where lateral films and cross-sectional imaging are rarely immediately available and management is often initiated prior to a formal radiology report.

A total of 200 positive cases (fractures) and 200 negative cases (non-fractures) from the PV dataset were randomly selected to form the reader study (MRMC) dataset. The sample size was chosen to balance the study requirements of as large a sample as possible with the logistic concerns of providing a dataset that the readers - all busy clinicians - would find feasible to evaluate. As was the case with the PV data, there was no overlap of patients with the Dev data and all patients were imaged from the emergency department. The balanced dataset (with equal numbers of fractures and non-fractures) was utilised to limit the number of cases for review by each reader.

External validation dataset

An international external validation (EV) dataset from Stanford University Hospital (California, USA) was obtained to assess the replicability of the performance of the model and its potential to generalise to new environments.

This dataset consisted of 93,455 images collected from patients at Stanford University Medical Centre who underwent a radiographic examination of the lower extremity between 2003 and 2014, as well as the associated exam reports(24). Each image was prospectively labeled as normal or abnormal by the attending radiologist at the time of initial interpretation. From this group, 46 positive and 100 negative hip radiographs were randomly selected. The negative images were reviewed by an attending radiologist to confirm "normal" labels or the presence of a fracture, and the positive (fracture) cases were confirmed either via follow-up radiographs with surgical fixation or review of follow-up cross-sectional imaging confirming the presence of a fracture.

Specific exclusions included cases with “burned-in” private health information (i.e., identifiable patient information stored within the image pixels themselves rather than in the metadata), and those cases which contained metalwork, resulting in a final EV dataset of 40 positive cases and 41 negative cases. 22 of the fractures (55%) involved the trochanteric region, and 18 of the fractures (45%) involved the femoral neck.

The data flow of cases and images in the RAH, Dev, PV, MRMC, and EV datasets is shown in Figure 1, with dataset characteristics in Table 1.

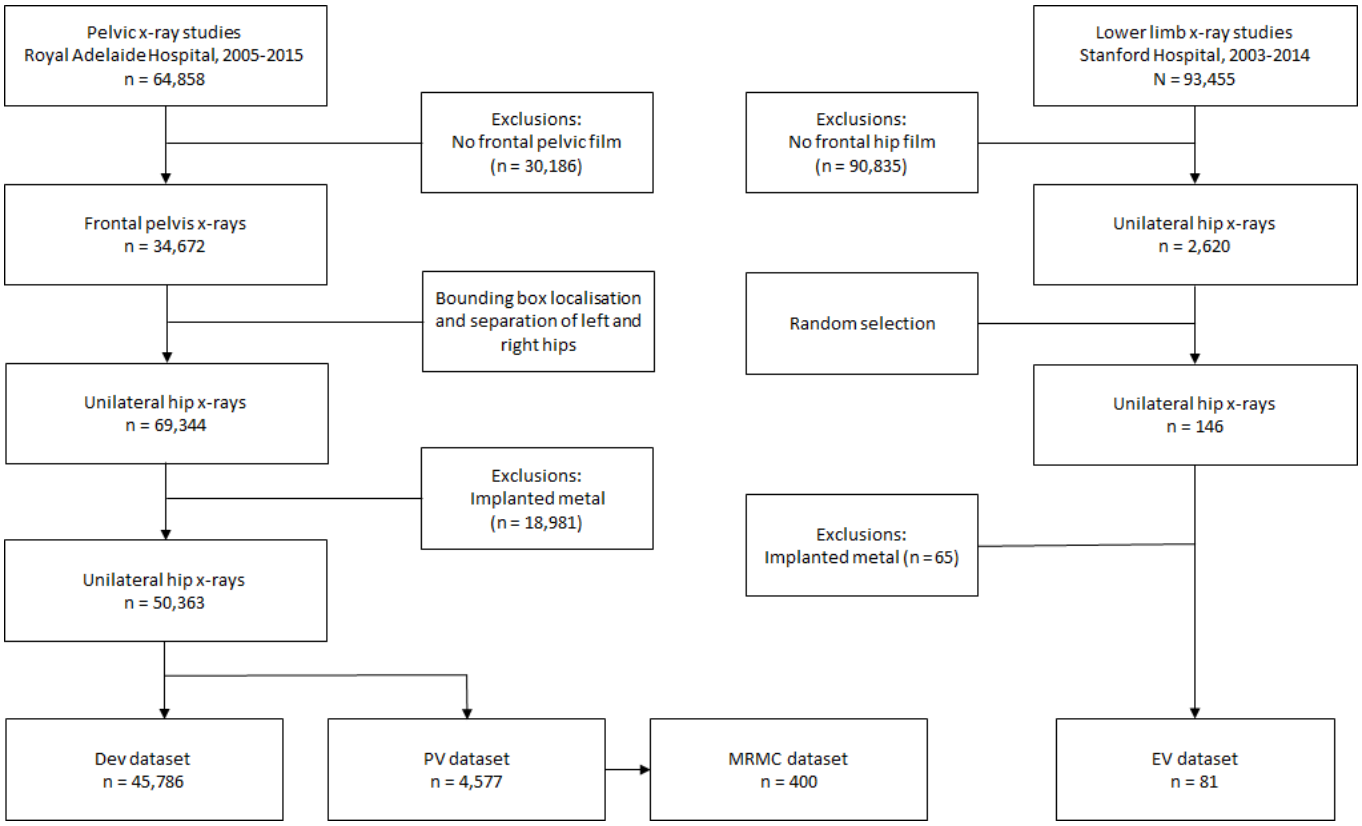


Figure 1: Data flow. The data acquisition process for the RAH and Stanford datasets. The RAH data is further divided into the Dev, PV, and MRMC datasets by randomisation at the patient level (i.e., no patients occur in both the Dev and PV/MRMC datasets).

Table 1: Dataset characteristics.

	Dev dataset	**PV dataset**	**MRMC dataset**	**EV dataset**
Patient characteristics				
Patients (n)	18,178	2,449	400	81
Frontal pelvic x-rays (n)	32,182	2,490	400	-
Unilateral hip images (n)	45,786	4,577	400	81
Mean age (SD) (years)	69.9 (22.0)	63.7 (25.4)	74.3 (24.0)	63.5 (23.5)
Sex (% female)	52%	48%	60%	61%
Ward of referral (% emergency department)	47%	100%	100%	N/A
Fracture prevalence (%)	11%	14%	50%	49%

Reader study

Thirteen practicing clinicians who might be expected to review these films in an emergency department setting were included in the reader study, with 5 radiologists in standard diagnostic conditions, as well as a mix of other clinicians (radiologists, surgical, emergency department and general practice doctors) who read the images under normal clinical conditions (i.e., without diagnostic quality monitors). In this context, “diagnostic conditions” refers to the use of

high-fidelity monitors and a fully-featured PACS viewer as required by the Royal Australian and New Zealand College of Radiologists(25) for all primary diagnostic reads performed by radiologists. “Clinical conditions” refers to the use of lower resolution monitors typically found in emergency departments and inpatient wards, typically used for case review or by non-radiologist clinicians. All readers reviewed the images with a locally developed web DICOM viewer, which provided a standard set of image manipulation tools such as windowing, zoom, and panning methods.

None of the readers had access to clinical information from the referral. While this is out of keeping with standard radiological practice, these preclinical experiments were designed to test visual performance at the task of proximal femoral fracture detection, rather than clinical diagnostic ability. The radiologists were only told that each case was an acute presentation to an emergency department, and the patient required pelvic x-ray imaging.

The five radiologists who were reporting in clinical conditions were all experienced at the task of musculoskeletal radiograph analysis, and consisted of 3 musculoskeletal specialists and 2 general radiologists. All radiologists were fully qualified consultant radiologists (i.e., were current Fellows of the Australian & New Zealand College of Radiologists) , and the musculoskeletal radiologists had completed appropriate subspecialty training. The radiologists had an average of 10.6 years of clinical experience (range = 5 to 19 years post fellowship). The radiologists were all recruited from a large, multi-site private radiology practice in South Australia (Dr Jones and Partners Pty Ltd).

Readers were asked to classify each x-ray into one of 4 categories: “definitely fracture”, “likely fracture, needs further imaging”, “likely not a fracture, needs further imaging”, or “no fracture”. These categories were dichotomized into ‘definite fracture’ (the first category) or ‘equivocal/non-fracture’ (latter three categories) for analysis, to estimate the potential of the model to avoid further follow-up imaging/investigation and therefore reduce delays to admission and surgery.

Primary analysis

The primary measure of performance for the AI algorithm and the readers was the area under the receiver operating characteristic curve (AUC) for the binary outcome fracture vs equivocal/non-fracture, and the primary comparison was between the AI algorithm and the 5 radiologists, including 3 subspecialised musculoskeletal (MSK) radiologists, who assessed the cases under diagnostic conditions. The results of the remaining readers (3 radiologists using non-diagnostic monitors and 5 non-radiologists) are presented for completeness and interest.

To estimate the “average” performance of the readers for comparison with the AI model, we adopted the well established and accepted practices of meta-analysis for diagnostic accuracy studies (20). By treating each reader as a distinct ‘diagnostic study’ with a known confusion matrix, we use summary receiver operating characteristic curve (SROC) analysis to summarise reader performance. As is the case in meta-analysis more broadly, this approach prevents the underestimation of human performance which is seen when sensitivity and specificity are independently pooled across readers(26,27), and allows for the robust statistical comparison of

AUC measures between test modalities. The 95% confidence intervals for the AI model were produced from a non-parametric bootstrap with 10,000 samples, and we performed null hypothesis testing on the difference of AUC measurements with the method reported by Delong(28).

Secondary analyses and external validation

We report multiple secondary findings to further characterise the performance of the AI model. These secondary findings are intended to be descriptive, demonstrating the specific diagnostic properties of the AI algorithm that are not captured in the summary performance of the primary analysis.

First, we show the performance of the AI algorithm at clinical prevalence, using the entire PV dataset of 4,577 unilateral hip x-rays, containing 640 fractures (the reader study was a subset of this larger PV dataset; the 200 fractures in the reader study were part of the 640 fractures in the test set).

Second, we report the sensitivity and specificity of the AI algorithm at an operating point selected on the basis of achieving the highest human sensitivity in the reader study (sens = 0.95). The operating point was selected by matching this level of performance on using the results from the development set by Gale et al(10) (i.e., not by matching this level of performance on the test set, which would be an information leak into the model from the held-out data) and corresponded with a threshold value of 0.62.

Third, we report the performance of the AI algorithm on an external validation dataset obtained from Stanford Hospital, reporting the AUC as well as the sensitivity and specificity at the selected operating point. The model was not retrained or fine-tuned prior to this assessment.

Fourth, we present the results of the full set of thirteen readers, including the non-radiologists and the radiologists who did not interpret the images under diagnostic conditions.

Algorithmic audit

We perform an algorithmic audit(29,30) to detect and characterise algorithmic errors, which we define as any outputs of the AI system which are inaccurate, including those which are inconsistent with the expected performance and those which can result in harm if undetected or detected too late. We followed the SMACTR framework of Raji et al(29) to identify sources of vulnerability to unexpected errors in the model and associated deployment environment. This process involved scoping and mapping the task, the model, and the environment, as well as defining the intended use and intended impact of the AI system. We then perform a Failure Mode and Effects Analysis (FMEA), and multiple subset analyses of the MRMC dataset including a Table 1 subset analysis, a Task-specific Subset analysis (TSSA), and an Exploratory Error Analysis (EEA). Similar to the other secondary analyses, these subset analyses are intended to be descriptive and null hypothesis significance testing has not been performed. The auditor is required to use their expertise to try to identify patterns in the errors (which have been called 'failure modes'). This process is qualitative, and intended to guide the audit and mitigation approach, rather than quantitative.

This methodology and structure of the medical algorithmic audit has been described in detail by Xiao et al (included in thesis).

The role of the funder

This research was unfunded, and there were no external interests involved in the data collection, analysis, interpretation, writing of the manuscript or the decision to submit.

Results

Reader study

In the primary performance comparison, the model AUC was 0.994 (95% CI = 0.988, 0.999), while the AUC of the SROC for the 5 radiologists was 0.969 (95% CI = 0.960, 0.978). The model ROC curve and radiologist SROC curve are shown in Figure 2.

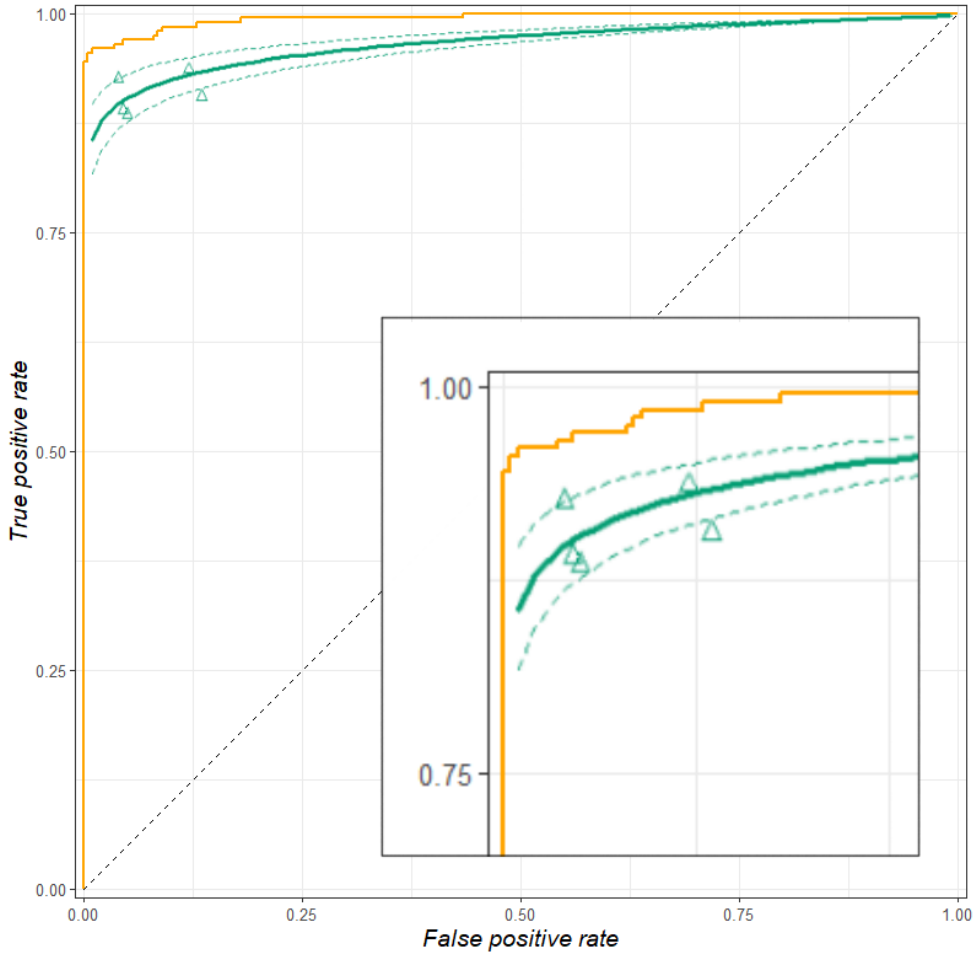


Figure 2: ROC results from the reader study, comparing the performance of the AI model (orange line) against the individual performance of the radiologists (green triangles) and the average human performance summarised with an SROC curve (solid green line) and confidence region (dotted green lines).

A confusion matrix demonstrating the number of false positive and false negative errors is presented in Extended Data Figure 1.

Results for a simulated “forced choice” experiment, where all definite or equivocal fracture responses from the readers were treated as a positive finding (instead of only the definite responses) is included in Extended Data Figure 2.

Secondary analyses

Full PV dataset results, external validation, and performance at the selected operating point

At the selected operating point, the sensitivity was 95.5 (95% CI = 91.8, 97.9) and the specificity was 99.5 (95% CI = 97.0, 100.0) on the MRMC dataset. These results were not significantly different from those found using the PV dataset (n = 4,577 images, 640 fractures) at clinical prevalence, where the AUC was 0.994 (p-value = 0.87), the sensitivity was 94.5 (95% CI = 91.7, 96.6) and the specificity was 99.1 (95% CI = 98.7, 99.4) at the selected operating point.

The model achieved an AUC of 0.98 (95% CI 0.93, 1.0) on the Stanford EV dataset, which was not significantly different (P=0.20) from the results reported on the Adelaide PV dataset. However, the operating point (of 0.62) shifted significantly when the model was applied to the Stanford validation set data, producing a sensitivity of 75.0 and a specificity of 100.0 (vs 95.5 and 99.5 respectively in the internal validation). In *post-hoc* analysis, the same sensitivity level (ie >95%) was achieved with an operating point of 0.0001, with a sensitivity of 97.4 and a specificity of 87.8.

Additional reader results

The performance of the additional readers (radiologists in non-diagnostic conditions and non-radiologist doctors) is shown in figure 3, both with plotting the sensitivity and specificity of individual readers, as well as summarising the performance of each group with SROC analysis.

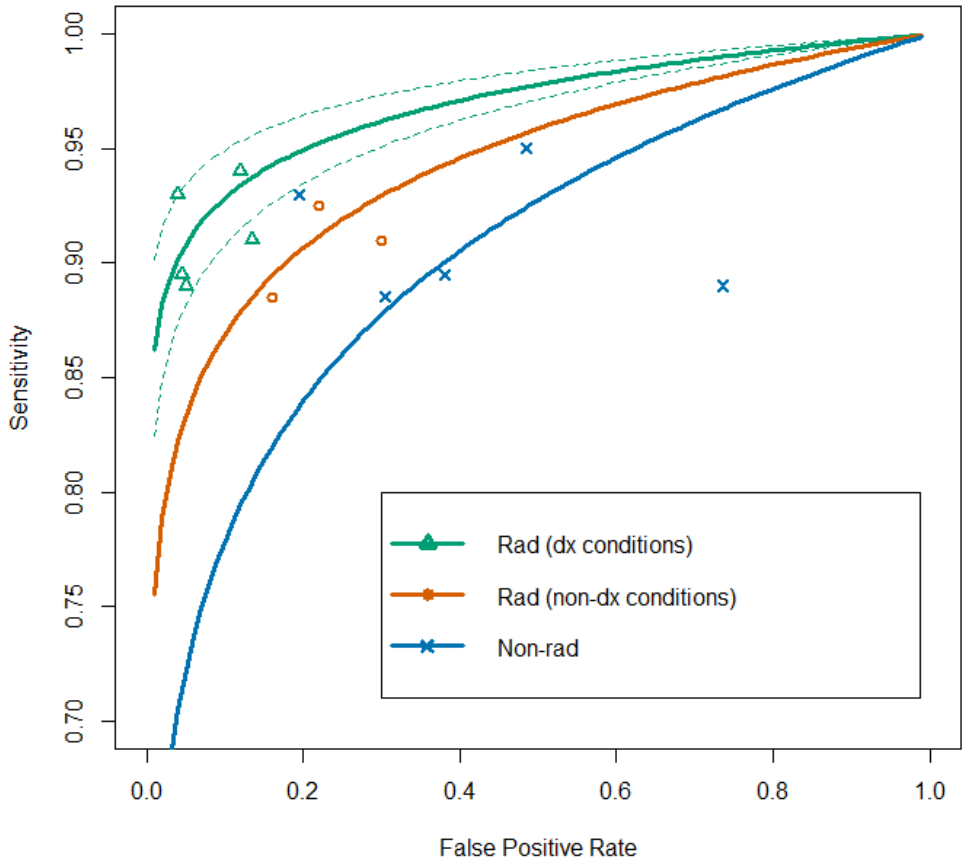


Figure 3: Additional performance results for other readers. We show results for the primary reader study radiologists in diagnostic conditions (n = 5, SROC-AUC = 0.969), as well as additional radiologists using non-diagnostic monitors (n = 3, SROC-AUC = 0.943), and non-radiologists (n=5, SROC-AUC = 0.902).

Algorithmic audit

The full set of audit artefacts are included in the Supplementary Information, including the full FMEA documents.

Subset performance

For task-specific subset analysis (TSSA), the fractures were labelled using a process of *schema completion(21)*, where an ontology of clinical relevant fracture subtypes was prospectively defined by a radiologist (LOR). These subtypes included features regarding the fracture location (i.e., subcapital, cervical, etc) and the fracture character (i.e., undisplaced, comminuted etc). To describe displacement we utilised the following system: "subtle" displacement = no or minimal cortical step, mild displacement = up to 1 cortical width, moderate displacement = up to half bone width, and severe displacement = more than half bone width. We did not distinguish between translation and angulation/tilt, but rather just referenced the most displaced component/region of the fracture. We chose to use this abbreviated descriptive system as we felt that it best described the useful elements of visual variation in the X-rays. Performance is reported in these subsets and compared against the performance of readers using ROC-AUC for the model and SROC-AUC for the readers.

The performance of the AI model for demographic subgroups (Table 1 Subgroup analysis) and clinically-relevant fracture subgroups (TSSA) are presented in Table 2 and Table 3.

Table 2: Subset analysis, demonstrating no aberrant model behaviour. While there is slightly lower performance in the oldest patient cohort (age > 80 yrs), a similar reduction in diagnostic accuracy is seen among the radiologists.

Subset (number of cases)	AUC (AI)	AUC (Rads)
Male (n = 160)	0.996	0.979
Female (n = 240)	0.994	0.967
Age < 40 (n = 53)	0.993	0.970
Age 40-60 (n = 63)	1.0	0.992
Age 60-80 (n = 59)	0.999	0.998
Age > 80 (n = 225)	0.988	0.967
Overall performance (n = 400)	**0.994**	**0.969**

Table 3: Task specific subset analysis, which reveals no obviously aberrant model behaviour. In particular, there is no large drop in model performance for intracapsular fractures (subcapital and cervical locations), which have distinct clinical implications.

Subset (number of cases)	AUC (AI)	AUC (Rads)
Subtle fractures (n = 9)	0.964	0.982
Mild displacement (n = 61)	0.998	0.969
Moderate displacement (n=56)	1.0	0.990
Severe displacement (n = 74)	1.0	0.946

Comminuted fracture (n = 75)	1.0	0.971
Subcapital location (n = 66)	0.999	0.980
Cervical location (n = 23)	0.984	0.982
Pertrochanteric location (n = 105)	0.999	0.958
Subtrochanteric location (n = 6)	0.970	0.968
Overall performance (n = 400)	**0.994**	**0.969**

Exploratory error analysis

The false positive and false negative cases were visually reviewed by a radiologist (LOR). Other than the subgroups already identified in the Table 1 subset analysis and TSSA, it was noted that cases with abnormal bone or joint appearances were overrepresented amongst the errors for the model. Targeted relabeling of the MRMC was undertaken, which revealed 6 cases with either abnormal trabecular patterns due to Paget's disease of the pelvis or femur, or severe femoral head deformities.

A subset analysis was performed, and while this was limited by the low number of cases involved, there was a large difference in the error rates for the overall MRMC dataset (error rate = 2.5%) and the cases with abnormal bones and joints (error rate = 50%). These cases are shown in Extended Data Figure 3.

No other obvious subsets were identified during EEA. There was a single further example of a surprising error; a false negative in a significantly displaced fracture. This case is shown in Extended Data Figure 4. The remaining 6 false negatives (i.e. those not already presented in Extended Data Figures 3 and 4) are displayed in Extended Data Figure 5.

It was noted on review of grad-cam saliency maps(31) (an interpretability method that produces a heat-map to visualise regions of the image that were most salient to the model decision) that the model had a tendency to focus on the inner cortex of the neck of femur region, which is part of a clinically relevant feature for proximal femoral fracture detection known as “Shenton’s line”(32). However, the saliency maps often did not highlight outer cortex fracture lines (Extended Data Figure 6), even when the model correctly diagnosed the fracture. In the example in Extended Data Figure 2, the outer cortex is clearly disrupted, but a plausible “intact” curve along Shenton’s line is able to be discerned (Extended Data Figure 2c). It is possible that this reflects a failure mode of the model: if displaced fracture elements form a “pseudo-Shenton’s line” like in this case, the model may misinterpret this to be a sign of an intact hip. This is speculative however, and little can be determined from this single error.

Discussion

We report results of a thorough investigation of a high-performance AI algorithm for the detection of proximal femoral fractures from frontal pelvic radiographs in emergency department patients. Overall the AI model achieved exceptional performance, outperforming radiologists in diagnostic reporting conditions on both the primary metric (AUC 0.994 vs 0.969) and by demonstrating both higher sensitivity and specificity than *any doctor* tested in the reader study. We also note that the performance was higher than the performance of a deep

learning model reported by Krogue et al (33) across their entire test set, perhaps due to the smaller dataset and image downsampling utilised in that research.

In response to concerns in the recent literature that pre-clinical AI testing can obscure various problems with AI models leading to an "implementation gap"(13) we performed a series of secondary analyses as well as an algorithmic audit. These concerns include: a lack of generalisability to different populations, unclear performance at true clinical prevalence (as the majority of reader studies are highly enriched), statistical limitations on the analytic comparison between human experts and AI models, and a lack of investigation into the unexpected behaviour of AI models within strata of a study population (where models with good performance can produce unacceptable answers in a subset of cases).

Apropos generalisability, our external validation results from a US cohort were informative. While the discriminative performance of the AI system (as measured by AUC) appears to be maintained, the drop in the sensitivity at the pre-specified operating point (95.5 to 75.0) would make the system clinically unusable in the new environment. While this could be mitigated by the selection of a new threshold, as shown when we demonstrate similar sensitivity and specificity in a post-hoc analysis (where the more minor drop in specificity simply reflects the small reduction in discriminative performance), this would require a localisation process to determine the new threshold in the new environment. We believe this is the first report of such behaviour in the medical AI literature, and we do not have a reasonable explanation; it appears at face value that any model that can maintain AUC across populations should also demonstrate

a fairly stable sensitivity and specificity at a given threshold. This issue requires further investigation.

We believe that using meta-analytic summary ROC analysis is more methodologically justified than alternatives, such as reporting the pooled sensitivity and specificity values of a group of doctors, and brings the field into conformity with broader biomedical practice. The SROC approach(20) solves the problem of underestimation of human performance when sensitivity and specificity are summarised independently, as well as better modelling the variation in performance across human readers.

Given the tendency of AI models to behave in unexpected ways (i.e., unlike a human expert would), the inclusion of an algorithmic audit appears to be worthwhile. As stated in Xiao et al (included in thesis), the audit approach changes the focus from '*what is the best performance this AI system can achieve*' to '*what is the worst mistake this AI system can make*'. Instead of simply reporting broad summary statistics, identifying what sort of cases the model fails on may assist in bridging the current gap between apparent high performance in preclinical testing and the uncertainty around the clinical implications of these models.

We note in particular that while the model demonstrates high performance, and does not appear to deviate from human performance in pre-specified subsets (Table 2 & 3), it does still make the occasional "inhuman" error, e.g., misdiagnosing a highly displaced fracture. We also note on saliency mapping that while the model reproduces some recognisable aspects of human practice (for example, the AI model appears to pay attention to Shenton's line), the visualisations nonetheless raise concerns about the regions *not* highlighted in the heatmaps. In

particular, the saliency maps almost never show strong activity along the outer region of the femoral neck, even in cases where the cortex in this area is clearly disrupted. While over-interpretation of saliency maps can be problematic due to known failings of these methods(34), these findings together raise the concern that, despite the model performing extremely well at the task of hip fracture detection when assessed with summary statistics, the model appears to be more "brittle" than a human. These results will hopefully be useful when planning to integrate the model into clinical workflows, and some possible strategies to mitigate various issues have been suggested in the algorithmic audit report (included in the Supplementary Information).

There are a number of limitations of this study. Firstly, we note that the model itself is limited, being unable to act on cases with implanted metalwork (albeit the system is able to automatically identify these cases and exclude them from analysis). Secondly, the sample size of the MRMC study was limited by the availability of readers; we determined a total dataset of 400 cases (200 positive and 200 negative cases) was as many as we could reasonably expect the readers to review, and only 5 radiologists reviewed the cases under diagnostic conditions as defined in the local standards of practice. We do note that the sample size compares well to other similar studies (11) and that the 95% confidence intervals are not excessively wide. Similarly, the sample size for the external validation is modest, but again the confidence intervals are reassuring from a clinical perspective. Thirdly, we recognise that despite the significant effort put into the algorithmic audit, there are important aspects of the model we could not test. Importantly, we were unable to access race and ethnicity data for our local population for subset testing, and despite performing an external validation on data from

Stanford Hospital we were unable to assess whether the deep learning model will generalise outside of the two clinical settings and the populations evaluated in this study.

Regarding the audit, we note that given the reliance on individual human interpretation and small underpowered subsets (or even individual examples) it would be reasonable to suspect that the findings of the audit and subset tests are not statistically reliable. We believe that such concerns are orthogonal to the purpose of the techniques, as the intention is to discover potential sources of unexpectedly poor clinical performance in a descriptive or exploratory manner, not to demonstrate a statistically robust effect or effect size.

Our study evaluated a high-performance proximal femoral fracture detection AI model, which outperforms highly trained clinical specialists in diagnostic conditions as well as other clinical readers in normal clinical environments. The performance of the AI system is maintained when applied to an external validation sample from an international site, and a thorough analysis of the behaviour of the AI system shows that it is mostly consistent with that of human experts. We also characterise the occasional aberrant or unexpected behaviour of the AI model to inform future clinical testing protocols.

We intend to test this model in a clinical environment, in the form of an interventional randomised control trial.

Acknowledgements

We would like to thank all of the clinicians who generously donated their time during the reader study: Dr M Moss, Dr T Kurmis, Dr C Tan, Dr J Heysen, Dr S Evans, Dr N Lavender, Dr N Bajic, Dr L E Yapp, Dr S Saha, Dr P Gribble, Dr S Mukhopadhaya, Dr M Cain, Dr SW Gan, Dr A Lovell, Dr G McCabe, and Dr R Luther.

Declaration of interests

GC, LJP and APB acknowledge the support received by the Australian Research Council's Discovery Projects funding scheme (project DP180103232). No other authors report any conflict of interest.

Author contributions

LOR, WG, GC, APB, and LJP conceived and planned the experiments. LOR gathered, cleaned, and labelled the primary validation data. All authors had access to the data. LOR and WG performed the experiments and analysis, and verified the data. TAB and MPL gathered and labelled the external validation dataset, and verified this data. LOR and LJP wrote the manuscript with critical feedback from all authors, and all authors were responsible for the decision to submit the manuscript.

Data sharing

The image data used to train and test the AI model is not shareable under the current agreement with the data custodian (SA Health).

The derived data, including the model and deidentified human reader classification outputs for the test data (as well as the related data dictionary), will be made available immediately following publication to anyone who wishes to access the data. Requests for access can be made to the corresponding author.

References

1. Brauer CA. Incidence and Mortality of Hip Fractures in the United States. JAMA. 2009;302(14):1573.
2. Kannus P, Parkkari J, Sievänen H, Heinonen A, Vuori I, Järvinen M. Epidemiology of hip fractures. Bone. 1996 Jan;18(1 Suppl):57S – 63S.
3. Dominguez S, Liu P, Roberts C, Mandell M, Richman PB. Prevalence of traumatic hip and pelvic fractures in patients with suspected hip fracture and negative initial standard radiographs--a study of emergency department patients. Acad Emerg Med. 2005 Apr;12(4):366–9.
4. Cannon J, Silvestri S, Munro M. Imaging choices in occult hip fracture. J Emerg Med. 2009 Aug;37(2):144–52.
5. Pincus D, Ravi B, Wasserstein D, Huang A, Paterson JM, Nathens AB, et al. Association Between Wait Time and 30-Day Mortality in Adults Undergoing Hip Fracture Surgery. JAMA. 2017 Nov 28;318(20):1994–2003.
6. Morrissey N, Iliopoulos E, Osmani AW, Newman K. Neck of femur fractures in the elderly: Does every hour to surgery count? Injury. 2017 Jun;48(6):1155–8.
7. Simunovic N, Devereaux PJ, Bhandari M. Surgery for hip fractures: Does surgical delay

affect outcomes? Indian J Orthop. 2011 Jan;45(1):27–32.

8. Shabat S, Heller E, Mann G, Gepstein R, Fredman B, Nyska M. Economic consequences of operative delay for hip fractures in a non-profit institution. Orthopedics. 2003 Dec;26(12):1197–9; discussion 1199.

9. Nagendran M, Chen Y, Lovejoy CA, Gordon AC, Komorowski M, Harvey H, et al. Artificial intelligence versus clinicians: systematic review of design, reporting standards, and claims of deep learning studies. BMJ. 2020 Mar 25;368:m689.

10. Gale W, Oakden-Rayner L, Carneiro G, Bradley AP, Palmer LJ. Detecting hip fractures with radiologist-level performance using deep neural networks. arXiv. 2017 Nov 17;

11. Liu X, Faes L, Kale AU, Wagner SK, Fu DJ, Bruynseels A, et al. A comparison of deep learning performance against health-care professionals in detecting diseases from medical imaging: a systematic review and meta-analysis. The Lancet Digital Health. 2019 Oct 1;1(6):e271–97.

12. Kohli A, Jha S. Why CAD Failed in Mammography. J Am Coll Radiol. 2018 Mar;15(3 Pt B):535–7.

13. Seneviratne MG, Shah NH, Chu L. Bridging the implementation gap of machine learning in healthcare. BMJ Innovations. 2020 Apr 1;6(2).

14. Lehman CD, Wellman RD, Buist DSM, Kerlikowske K, Tosteson ANA, Miglioretti DL, et al. Diagnostic Accuracy of Digital Screening Mammography With and Without Computer-Aided Detection. JAMA Intern Med. 2015 Nov;175(11):1828–37.

15. Lyell D, Coiera E. Automation bias and verification complexity: a systematic review. J Am Med Inform Assoc. 2017 Mar 1;24(2):423–31.

16. Zech JR, Badgeley MA, Liu M, Costa AB, Titano JJ, Oermann EK. Variable generalization performance of a deep learning model to detect pneumonia in chest radiographs: A cross-sectional study. PLoS Med. 2018 Nov;15(11):e1002683.

17. Nam JG, Park S, Hwang EJ, Lee JH, Jin K-N, Lim KY, et al. Development and Validation of Deep Learning–based Automatic Detection Algorithm for Malignant Pulmonary Nodules on Chest Radiographs. Radiology. 2019;290(1):218–28.

18. McKinney SM, Sieniek M, Godbole V, Godwin J, Antropova N, Ashrafian H, et al. International evaluation of an AI system for breast cancer screening. Nature. 2020 Jan;577(7788):89–94.

19. Pooch EHP, Ballester P, Barros RC. Can We Trust Deep Learning Based Diagnosis? The Impact of Domain Shift in Chest Radiograph Classification. Thoracic Image Analysis. 2020;74–83.

20. Oakden-Rayner L, Palmer L. Docs are ROCs: A simple off-the-shelf approach for estimating average human performance in diagnostic studies. arXiv. 2020 Sep 23;

21. Oakden-Rayner L, Dunnmon J, Carneiro G, Re C. Hidden stratification causes clinically meaningful failures in machine learning for medical imaging. In: Proceedings of the ACM Conference on Health, Inference, and Learning. New York, NY, USA: Association for Computing Machinery; 2020. p. 151–9. (CHIL '20).

22. Raji ID, Smart A, White RN, Mitchell M, Gebru T, Hutchinson B, et al. Closing the AI accountability gap: defining an end-to-end framework for internal algorithmic auditing. In: Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency. New York, NY, USA: Association for Computing Machinery; 2020. p. 33–44. (FAT* '20).

23. Huang G, Liu Z, Van Der Maaten L, Weinberger KQ. Densely connected convolutional networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition. 2017. p. 4700–8.

24. Varma M, Lu M, Gardner R, Dunnmon J, Khandwala N, Rajpurkar P, et al. Automated abnormality detection in lower extremity radiographs using deep learning. Nature Machine Intelligence. 2019 Dec 1;1(12):578–83.

25. The Royal Australian and New Zealand College of Radiologists. 3.6 Computers and Automated Equipment. Standards of Practice for Diagnostic and Interventional Radiology, Version 102 [Internet]. 2020 [cited 2020 Sep 3]. Available from: https://www.ranzcr.com/documents/510-ranzcr-standards-of-practice-for-diagnostic-and-interventional-radiology/file

26. Irwig L, Tosteson AN, Gatsonis C, Lau J, Colditz G, Chalmers TC, et al. Guidelines for meta-analyses evaluating diagnostic tests. Ann Intern Med. 1994 Apr 15;120(8):667–76.

27. Gatsonis C, Paliwal P. Meta-analysis of diagnostic and screening test accuracy evaluations: methodologic primer. AJR Am J Roentgenol. 2006 Aug;187(2):271–81.

28. DeLong ER, DeLong DM, Clarke-Pearson DL. Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. Biometrics. 1988 Sep;44(3):837–45.

29. Raji ID, Smart A, White RN, Mitchell M, Gebru T, Hutchinson B, et al. Closing the AI Accountability Gap: Defining an End-to-End Framework for Internal Algorithmic Auditing. arXiv. 2020 Jan 3;

30. Mahajan V, Venugopal VK, Murugavel M, Mahajan H. The Algorithmic Audit: Working with Vendors to Validate Radiology-AI Algorithms-How We Do It. Acad Radiol. 2020 Jan;27(1):132–5.

31. Selvaraju RR, Cogswell M, Das A, Vedantam R, Parikh D, Batra D. Grad-cam: Visual

explanations from deep networks via gradient-based localization. In: Proceedings of the IEEE international conference on computer vision. 2017. p. 618–26.

32. Jones DHA. Shenton's line. J Bone Joint Surg Br. 2010 Sep;92(9):1312–5.

33. Krogue JD, Cheng KV, Hwang KM, Toogood P, Meinberg EG, Geiger EJ, et al. Automatic Hip Fracture Identification and Functional Subclassification with Deep Learning. Radiol Artif Intell. 2020 Mar;2(2):e190023.

34. Adebayo J, Gilmer J, Muelly M, Goodfellow I, Hardt M, Kim B. Sanity Checks for Saliency Maps. In: Bengio S, Wallach H, Larochelle H, Grauman K, Cesa-Bianchi N, Garnett R, editors. Advances in Neural Information Processing Systems 31. Curran Associates, Inc.; 2018. p. 9505–15.

Conclusions

The implementation gap remains the largest barrier to safe and efficacious medical AI deployment. Given the current *status quo*, where AI models are being cleared by regulators and sold into clinics *without* pre-market evidence in the form of clinical trials, mechanisms to close the implementation gap are of critical importance and must be considered.

AI researchers tend towards technological solutionism, consistent with the origin of the field in computer science. Where fully automated (and therefore near frictionless) methods are favoured, such solutions do not yet exist and AI models are already in use in clinics, being used on real patients. I have therefore focussed on approaches that utilise the only resources that can currently be applied to the implementation gap - humans and their expertise. While exhaustive labelling, subset testing, and algorithmic auditing are time consuming, they are the only reliable methods currently available to address the shortcomings of current testing procedures.

The critical role of human expertise and oversight has been recognised in AI applications more broadly. Safety in medical AI has analogues in other applications which involve the legitimate risk of harm, including self-driving vehicles and policing/judicial algorithms. Major sources of harm have only been uncovered with thoughtful human analysis and effort [44–46]. The similarity between the incautious use of AI algorithms in these domains and the current medical AI regulatory environment is concerning, and it seems reasonable to assume that there is a similar risk of serious harm.

Indeed, one major limitation of this thesis can be appreciated when looking at these other domains; this work does not directly address a major source of potential harm in medical AI, that of racial and gender bias in medical AI models. AI healthcare disparities, which may operate to bias *both* the input data and the data labels, can be considered a form of hidden stratification, where harm is caused by underperformance among specific patient subgroups. It is clear, however, that both the prevalence of these disparities and the indefensibility of failing to perform targeted testing for these patient populations (where demographic information is often available at no additional cost) is of sufficient concern and urgency to require specific intervention beyond the recommendations contained in my thesis. There are numerous examples of medical AI models which replicate or even exacerbate the sociocultural disparities that already exist in medical practice [14,47], and work in this space to identify and mitigate the harms posed by this technology to under-served patients requires far more visibility.

Similarly, the complex interactions between human users and AI models is poorly understood but is recognised as a further factor contributing to the implementation gap. Automation bias, laboratory effects, and false alarm fatigue have all been cited as potential sources of poor real-world performance in medical algorithms [10,48,49]. Much of the literature on AI human-computer interaction is confusing and contradictory, suggesting that a great deal of further work is required if we truly intend to close the implementation gap for systems which keep humans in-the-loop.

I note that my thesis does not reflect a desire to avoid clinical testing of AI algorithms. Instead, I believe that clinical testing is critical to AI safety in healthcare. This work simply recognises the low likelihood of widespread clinical testing in the current healthcare and regulatory environments and proposes a series of achievable techniques to reduce harms given that context.

References

1. Liu, X. *et al.* A comparison of deep learning performance against health-care professionals in detecting diseases from medical imaging: a systematic review and meta-analysis. *The Lancet Digital Health* **1**, e271–e297 (2019).
2. Nagendran, M. *et al.* Artificial intelligence versus clinicians: systematic review of design, reporting standards, and claims of deep learning studies. *BMJ* **368**, m689 (2020).
3. Muehlematter, U. J., Daniore, P. & Vokinger, K. N. Approval of artificial intelligence and machine learning-based medical devices in the USA and Europe (2015–20): a comparative analysis. *The Lancet Digital Health* **3**, e195–e203 (2021).
4. Harvey, H. B. & Gowda, V. How the FDA regulates AI. *Acad. Radiol.* **27**, 58–61 (2020).
5. Benjamens, S., Dhunnoo, P. & Meskó, B. The state of artificial intelligence-based FDA-approved medical devices and algorithms: an online database. *NPJ Digit Med* **3**, 118 (2020).
6. Center for Devices & Radiological Health. Software as a Medical Device (SAMD): Clinical Evaluation - Guidance. https://www.fda.gov/regulatory-information/search-fda-guidance-documents/software-medical-device-samd-clinical-evaluation.
7. Office of the Commissioner. FDA Releases Artificial Intelligence/Machine Learning Action Plan. https://www.fda.gov/news-events/press-announcements/fda-releases-artificial-intelligencemachine-learning-action-plan (2021).
8. Seneviratne, M. G., Shah, N. H. & Chu, L. Bridging the implementation gap of machine learning in healthcare. *BMJ Innovations* **6**, (2020).
9. Lehman, C. D. *et al.* Diagnostic Accuracy of Digital Screening Mammography With and Without Computer-Aided Detection. *JAMA Intern. Med.* **175**, 1828–1837 (2015).
10. Kohli, A. & Jha, S. Why CAD Failed in Mammography. *J. Am. Coll. Radiol.* **15**, 535–537 (2018).
11. Zech, J. R. *et al.* Variable generalization performance of a deep learning model to detect

pneumonia in chest radiographs: A cross-sectional study. *PLoS Med.* **15**, e1002683 (2018).

12. Badgeley, M. A. *et al.* Deep learning predicts hip fracture using confounding patient and healthcare variables. *NPJ Digit Med* **2**, 31 (2019).
13. Winkler, J. K. *et al.* Association Between Surgical Skin Markings in Dermoscopic Images and Diagnostic Performance of a Deep Learning Convolutional Neural Network for Melanoma Recognition. *JAMA Dermatol.* (2019) doi:10.1001/jamadermatol.2019.1735.
14. Obermeyer, Z., Powers, B., Vogeli, C. & Mullainathan, S. Dissecting racial bias in an algorithm used to manage the health of populations. *Science* **366**, 447–453 (2019).
15. Seyyed-Kalantari, L., Liu, G., McDermott, M., Chen, I. Y. & Ghassemi, M. CheXclusion: Fairness gaps in deep chest X-ray classifiers. *arXiv [cs.CV]* (2020).
16. Park, Y. *et al.* Evaluating artificial intelligence in medicine: phases of clinical research. *JAMIA Open* **3**, 326–331 (2020).
17. Rivera, S. C. *et al.* Guidelines for clinical trial protocols for interventions involving artificial intelligence: the SPIRIT-AI Extension. *BMJ* **370**, m3210 (2020).
18. Liu, X. *et al.* Reporting guidelines for clinical trial reports for interventions involving artificial intelligence: the CONSORT-AI Extension. *BMJ* **370**, m3164 (2020).
19. a Medical Device (SaMD) Working Group, I. S. as. 'Software as a Medical Device': Possible Framework for Risk Categorization and Corresponding Considerations. in (International Medical Device Regulators Forum, 2014).
20. Aggarwal, R. *et al.* Diagnostic accuracy of deep learning in medical imaging: a systematic review and meta-analysis. *npj Digital Medicine* **4**, 1–23 (2021).
21. Gebru, T. *et al.* Datasheets for Datasets. *arXiv [cs.DB]* (2018).
22. Sun, C., Shrivastava, A., Singh, S. & Gupta, A. Revisiting unreasonable effectiveness of data in deep learning era. in *Proceedings of the IEEE international conference on computer vision* 843–852 (2017).
23. BreastScreen Australia monitoring report 2020.

https://www.aihw.gov.au/reports/cancer-screening/breastscreen-australia-monitoring-report-2020/contents/summary.

24. Rolnick, D., Veit, A., Belongie, S. & Shavit, N. Deep Learning is Robust to Massive Label Noise. *arXiv [cs.LG]* (2017).
25. Algan, G. & Ulusoy, İ. Label Noise Types and Their Effects on Deep Learning. *arXiv [cs.CV]* (2020).
26. Wang, X. *et al.* Chestx-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. in *Proceedings of the IEEE conference on computer vision and pattern recognition* 2097–2106 (2017).
27. Irvin, J. *et al.* Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison. in *Proceedings of the AAAI Conference on Artificial Intelligence* vol. 33 590–597 (2019).
28. Johnson, A. E. W. *et al.* MIMIC-CXR, a de-identified publicly available database of chest radiographs with free-text reports. *Sci Data* **6**, 317 (2019).
29. Oakden-Rayner, L. Exploring Large-scale Public Medical Image Datasets. *Acad. Radiol.* **27**, 106–112 (2020).
30. Rajpurkar, P. *et al.* MURA: Large Dataset for Abnormality Detection in Musculoskeletal Radiographs. *arXiv [physics.med-ph]* (2017).
31. Oakden-Rayner, L., Dunnmon, J., Carneiro, G. & Re, C. Hidden stratification causes clinically meaningful failures in machine learning for medical imaging. in *Proceedings of the ACM Conference on Health, Inference, and Learning* 151–159 (Association for Computing Machinery, 2020).
32. Tariq, A. *et al.* Current Clinical Applications of Artificial Intelligence in Radiology and Their Best Supporting Evidence. *J. Am. Coll. Radiol.* **17**, 1371–1381 (2020).
33. Geirhos, R. *et al.* Shortcut Learning in Deep Neural Networks. *arXiv [cs.CV]* (2020).
34. Center for Devices & Radiological Health. Clinical Performance Assessment: Considerations for CAD Devices.

https://www.fda.gov/regulatory-information/search-fda-guidance-documents/clinical-performance-assessment-considerations-computer-assisted-detection-devices-applied-radiology.

35. Cohen, J. F. *et al.* STARD 2015 guidelines for reporting diagnostic accuracy studies: explanation and elaboration. *BMJ Open* **6**, e012799 (2016).
36. Obuchowski, N. A. Receiver operating characteristic curves and their use in radiology. *Radiology* **229**, 3–8 (2003).
37. Obuchowski, N. A. ROC analysis. *AJR Am. J. Roentgenol.* **184**, 364–372 (2005).
38. Dendumrongsup, T. *et al.* Multi-reader multi-case studies using the area under the receiver operator characteristic curve as a measure of diagnostic accuracy: systematic review with a focus on quality of data reporting. *PLoS One* **9**, e116018 (2014).
39. Oakden-Rayner, L. & Palmer, L. Docs are ROCs: A simple off-the-shelf approach for estimating average human performance in diagnostic studies. *arXiv [stat.ME]* (2020).
40. D'Amour, A. *et al.* Underspecification Presents Challenges for Credibility in Modern Machine Learning. *arXiv [cs.LG]* (2020).
41. Ribeiro, M. T., Singh, S. & Guestrin, C. 'Why Should I Trust You?': Explaining the Predictions of Any Classifier. in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* 1135–1144 (Association for Computing Machinery, 2016).
42. Bussone, A., Stumpf, S. & O'Sullivan, D. The Role of Explanations on Trust and Reliance in Clinical Decision Support Systems. in *2015 International Conference on Healthcare Informatics* 160–169 (2015).
43. Selbst, A. D. & Barocas, S. The intuitive appeal of explainable machines. *Fordham Law Rev.* (2018).
44. Using Artificial Intelligence and Algorithms. https://www.ftc.gov/news-events/blogs/business-blog/2020/04/using-artificial-intelligence-algorithms (2020).
45. Ghassemi, M., Pushkarna, M., Wexler, J., Johnson, J. & Varghese, P. ClinicalVis: Supporting

Clinical Task-Focused Design Evaluation. *arXiv [cs.HC]* (2018).

46. Eiband, M., Buschek, D., Kremer, A. & Hussmann, H. The Impact of Placebic Explanations on Trust in Intelligent Systems. in *Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems* 1–6 (Association for Computing Machinery, 2019).
47. Ghassemi, M., Oakden-Rayner, L. & Beam, A. L. The false hope of current approaches to explainable artificial intelligence in health care. *Lancet Digit Health* **3**, e745–e750 (2021).
48. Angwin, J., Larson, J., Kirchner, L. & Mattu, S. Machine Bias: There's software used across the country to predict future criminals. And it's biased against blacks. *ProPublica* https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing.
49. Buolamwini, J. & Gebru, T. Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification. in *Proceedings of the 1st Conference on Fairness, Accountability and Transparency* (eds. Friedler, S. A. & Wilson, C.) vol. 81 77–91 (PMLR, 2018).
50. Dressel, J. & Farid, H. The accuracy, fairness, and limits of predicting recidivism. *Sci Adv* **4**, eaao5580 (2018).
51. Sarkar, R. *et al.* Performance of intensive care unit severity scoring systems across different ethnicities in the USA: a retrospective observational study. *Lancet Digit Health* **3**, e241–e249 (2021).
52. Gur, D. *et al.* The 'Laboratory' Effect: Comparing Radiologists' Performance and Variability during Prospective Clinical and Laboratory Mammography Interpretations. *Radiology* **249**, 47–53 (2008).
53. Lyell, D. & Coiera, E. Automation bias and verification complexity: a systematic review. *J. Am. Med. Inform. Assoc.* **24**, 423–431 (2017).