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

Lauren Oakden-Rayner, William Gale, Thomas A Bonham, Matthew P Lungren, Gustavo Carneiro, Andrew P Bradley, Lyle J Palmer

Summary

Background Proximal femoral fractures are an important clinical and public health issue associated with substantial morbidity and early mortality. Artificial intelligence might offer improved diagnostic accuracy for these fractures, but typical approaches to testing of artificial intelligence models can underestimate the risks of artificial intelligence-based diagnostic systems.

Methods We present a preclinical evaluation of a deep learning model intended to detect proximal femoral fractures in frontal x-ray films in emergency department patients, trained on films from the Royal Adelaide Hospital (Adelaide, SA, Australia). This evaluation included a reader study comparing the performance of the model against five radiologists (three musculoskeletal specialists and two general radiologists) on a dataset of 200 fracture cases and 200 non-fractures (also from the Royal Adelaide Hospital), an external validation study using a dataset obtained from Stanford University Medical Center, CA, USA, and an algorithmic audit to detect any unusual or unexpected model behaviour.

Findings In the reader study, the area under the receiver operating characteristic curve (AUC) for the performance of the deep learning model was 0·994 (95% CI 0·988–0·999) compared with an AUC of 0·969 (0·960–0·978) for the five radiologists. This strong model performance was maintained on external validation, with an AUC of 0·980 (0·931–1·000). However, the preclinical evaluation identified barriers to safe deployment, including a substantial shift in the model operating point on external validation and an increased error rate on cases with abnormal bones (eg, Paget's disease).

Interpretation The model outperformed the radiologists tested and maintained performance on external validation, but showed several unexpected limitations during further testing. Thorough preclinical evaluation of artificial intelligence models, 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 None.

Lancet Digit Health 2022

Published Online
April 5, 2022
https://doi.org/10.1016/S2589-7500(22)00004-8

See Online/Viewpoint
https://doi.org/10.1016/S2589-7500(22)00003-6

School of Public Health (L Oakden-Rayner MBBS, Prof L J Palmer PhD), **Australian Institute for Machine Learning** (L Oakden-Rayner, W Gale BSc, Prof G Carneiro PhD, Prof L J Palmer), **and School of Computer Science** (W Gale), **University of Adelaide, Adelaide, SA, Australia; Stanford University School of Medicine, Department of Radiology, Stanford, CA, USA** (T A Bonham BS, M P Lungren MD); **Stanford Artificial Intelligence in Medicine and Imaging Center, Stanford University, Stanford, CA, USA** (M P Lungren); **Science and Engineering Faculty, Queensland University of Technology, Brisbane, QLD, Australia** (Prof A P Bradley PhD)

Correspondence to:
Dr Lauren Oakden-Rayner, Australian Institute for Machine Learning, University of Adelaide, Adelaide, SA 5000, Australia
lauren.oakden-rayner@adelaide.edu.au

Introduction

Hip fractures are an important global clinical and public health issue. In older people, proximal femoral fractures are the second most frequent cause of hospitalisation and are 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, primarily due to the subtle findings in a subset of these fractures.[3] Of those patients undergoing additional imaging, only around a third are ultimately ever diagnosed with a fracture.[3,4] Not only does further imaging increase the diagnostic costs, the burden on doctors and patients, and resource use, but so-called occult fractures could also lead to delayed diagnoses and concomitant worse patient outcomes, including increased mortality rate,[5,6] length of hospitalisation,[7] and cost of care.[8]

Improved diagnostic accuracy at first clinical presentation could plausibly reduce both harms and costs. Many studies have reported that artificial intelligence systems might exceed human performance for certain diagnostic tasks.[9] To reduce the rates of misdiagnosis or incomplete diagnosis of an initial radiograph in an emergency department, we have previously developed a deep learning-based proximal femoral fracture detection model with exceptional performance characteristics.[10] We evaluate the performance of the deep learning model, and compare this against the current standard of care (clinical radiologists) in a multi-reader, multi-case (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 done to show how these results translate into clinical practice.[9] Historically, computer-aided diagnosis systems have often performed unexpectedly poorly in the clinical setting

Research in context

Evidence before this study

We searched Google Scholar on Dec 10, 2019, for literature published up to Dec 10, 2019, with no language restrictions, on: deep learning-based detection of hip fractures with use of the keywords "hip fracture" or "proximal femoral fracture", and "deep learning" or "artificial intelligence"; and algorithmic audits of deep learning studies using the keywords "deep learning" or "artificial intelligence" and "audit". These literature searches were repeated on Dec 1, 2021. The literature on hip fracture detection using deep learning models was scarce and only six relevant studies were retrieved before study inception on Dec 10, 2019. The majority of studies reported internal performance of the artificial intelligence model only, with few reader studies and little external validation. There were no studies reporting further analysis into unexpected model behaviour or failure modes. Furthermore, to our knowledge, no audits of medical artificial intelligence systems have been reported.

Added value of this study

This study presents a thorough preclinical evaluation of a medical artificial intelligence system (trained to detect proximal femoral fractures on plain film imaging). Despite high performance of the model, which outperformed human experts in the task of proximal femoral fracture detection, an evaluation including algorithmic auditing showed unexpected and potentially harmful algorithmic behaviour.

Implications of all the available evidence

Thorough evaluation of artificial intelligence systems, including algorithmic auditing, can identify barriers to safe artificial intelligence deployment that might not be appreciated during standard preclinical testing and which could cause significant harm. Regulators, medical governance bodies, and professional groups should consider the need for more comprehensive preclinical testing of artificial intelligence before clinical deployment.

despite promising preclinical evaluations,[12] a concept known as the implementation gap.[13] Factors posited to explain poor clinical performance include: the misapplication of models outside of intended use cases,[14,15] a varying ability to generalise to new clinical environments,[16–19] statistical flaws when estimating the pooled performance and variability of human readers,[20] and the occurrence of unidentified poor performance in clinically important subgroups of cases.[21] Few preclinical artificial intelligence research studies have addressed these concerns; for example, external validation—an assessment of the ability of a model to generalise to new environments—has only been done in around a third of studies.[11] In the past few years, formal algorithmic auditing has been proposed[22] as a mechanism to identify and mitigate sources of undesirable machine learning model behaviour.

We performed a preclinical evaluation of a previously developed high-performance proximal femoral fracture model.[10] This work is intended to reflect current best practice for preclinical assessment, by meeting the following criteria: an MRMC study design that is adequately powered to determine the relative performance of an artificial intelligence model and human experts; external validation of a model on international data to attempt to replicate the results and identify any challenges for generalisation to new clinical environments; and an algorithmic audit to identify any unexpected behaviour of a deep learning model and to estimate any gaps between preclinical performance and safe 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 that had no patient overlap with the study datasets, consisting of 45 786 unilateral proximal femoral x-ray images with a fracture prevalence of 11% (4861 fractures).

A large local dataset was obtained from the Royal Adelaide Hospital (Adelaide, SA, Australia), a tertiary public teaching hospital that serves adult patients (aged >16 years). The Royal Adelaide Hospital dataset included all frontal pelvic x-rays ordered between Jan 1, 2005, and Dec 31, 2015, as part of standard clinical care, obtained using a wide variety of x-ray equipment and imaging techniques (the exact models of scanning equipment and imaging parameters used were not available in our dataset). X-ray studies with no frontal pelvic film were excluded. Likewise, cases with previous surgical intervention (implanted metalwork) were excluded because fractures in postoperative hips were thought to represent a visually and clinically distinct class of injury, and this work focused only on the detection of fractures in preoperative hips.

Visual assessments to exclude cases were performed by a series of so-called helper artificial intelligence models developed and validated during earlier work, with human review of all included films to ensure the validity of inclusion.[10]

Primary validation dataset

The primary validation dataset was randomly selected (at the patient level) from the emergency department cases in the Royal Adelaide Hospital dataset. Emergency department referrals were used for the primary validation dataset because we considered this to be the most clinically challenging setting; lateral films and cross-sectional imaging are often not immediately available

and management is often initiated before a formal radiology report. A total of 4577 unilateral hip x-rays were selected, including 640 proximal femoral fractures. The ground truth for fracture status was determined through a combination of x-ray reports, follow-up imaging with CT (91 patients) or MRI scans (five patients), and surgical records, with a follow-up period of at least 6 months. Mortality records were searched but identified no further cases of proximal femoral fractures. The majority of proximal femoral fracture cases were surgically validated (585 fractures, 91·4%), meaning the patients were surgically treated for fracture. The rest of the patients either did not receive surgery (ie, they died before surgery or were palliated), or they were transferred to other institutions before treatment.

The primary validation dataset was intended to investigate the application of our model to de-novo clinical cases; such cases were not available to the model during training. The remainder of the Royal Adelaide Hospital dataset (45 786 images) was used for model development.

MRMC dataset

200 positive cases (fractures) and 200 negative cases (non-fractures) from the primary validation dataset were randomly selected to form the reader study (MRMC) dataset. The sample size was chosen to balance the requirements for as large a sample as possible with the need to provide a dataset that busy clinicians would find feasible to read. There was no overlap of patients with the development dataset and all patients were imaged from the emergency department.

External validation dataset

An international external validation dataset from Stanford University Medical Center (Stanford, CA, USA) was obtained to assess the replicability and generalisability of our artificial intelligence model.

The external validation dataset consisted of 93 455 images collected from patients at Stanford University Medical Center who underwent a radiographic examination of the lower extremity between Jan 1, 2003, and Dec 31, 2014, as well as the associated examination reports.[24] Each image was prospectively labelled as normal or abnormal by the attending radiologist at first presentation. From this group, 46 positive and 100 negative hip radiographs were randomly selected. The negative images were reviewed by an attending radiologist to exclude 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.

Exclusion of images with burned-in private health information (ie, identifiable information stored within the image pixels themselves) and those which contained metalwork resulted in a final external validation dataset of 40 positive cases and 41 negative cases. Among the positive cases, 22 (55%) involved fractures in the trochanteric region, and 18 (45%) involved fractures of the femoral neck.

Reader study

13 practising clinicians who might be expected to review these films in an emergency department setting were included in the reader study, with five radiologists in standard diagnostic conditions, as well as a mix of eight other clinicians (radiologists and surgical, emergency department, and general practice doctors) who read the images under normal clinical conditions (ie, without diagnostic quality monitors). In this context, diagnostic conditions refers to the use of high-fidelity monitors and a fully featured Picture Archiving and Comunication System 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, commonly used for case review or by non-radiologist clinicians. All readers reviewed the images with a locally developed web Digital Imaging and Communications in Medicine 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. 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 diagnostic conditions consisted of three musculoskeletal specialists and two general radiologists. All radiologists were fully qualified, and the musculoskeletal radiologists had completed appropriate subspecialty training. The radiologists had a median of 10 years of clinical experience (range of 5–19 years post fellowship).

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

Outcomes

The primary measure of performance for the deep learning algorithm and the readers was the area under the receiver operating characteristic (ROC) curve (AUC) for the binary outcome, and the primary comparison was between the algorithm and the five radiologists reading under diagnostic conditions.

To estimate the average performance of the readers, we adopted the well established practice of meta-analysis for diagnostic accuracy studies.[20] By treating each reader as a

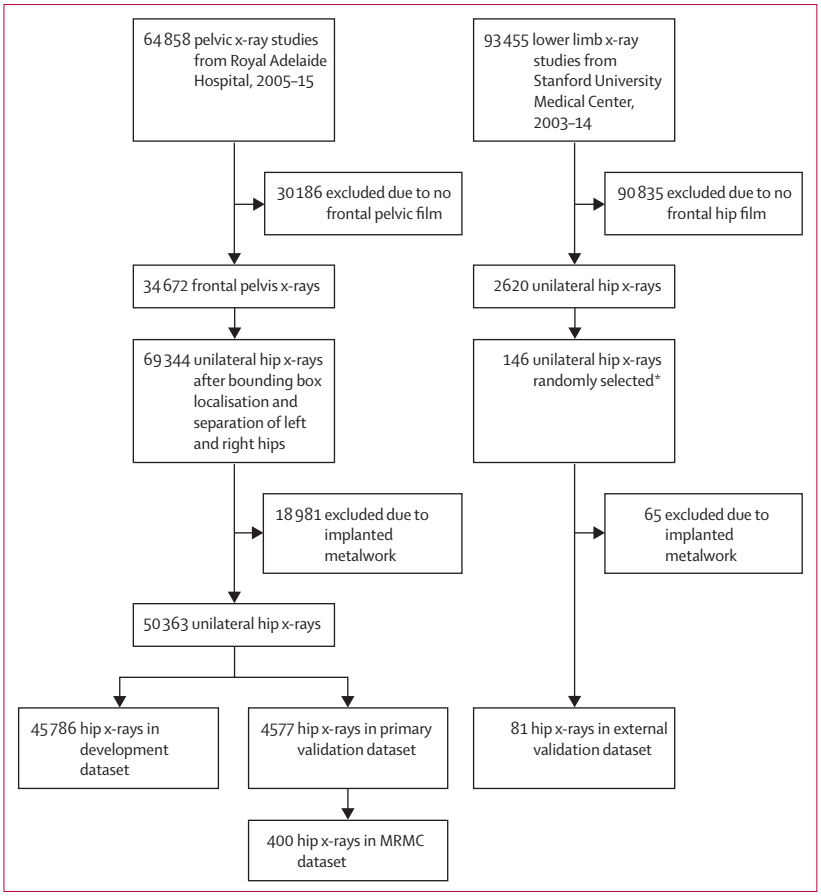

Figure 1: **Study flow diagram**
The Royal Adelaide Hospital data were further divided into the development, primary validation, and MRMC datasets by randomisation at the patient level (ie, no patients occur in both the development and primary validation or MRMC datasets). MRMC=multi-reader, multi-case. *Only x-rays without burned-in private health information were selected.

distinct diagnostic study with a known confusion matrix, we used summary ROC curve (SROC) analysis to summarise reader performance. This approach prevents the underestimation of human performance that is seen when sensitivity and specificity are independently pooled across readers,[26,27] and allows for the robust statistical comparison of AUC measures between human and artificial intelligence decision makers. The 95% CIs for the deep learning 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 and colleagues.[28]

See Online for appendix

We report multiple secondary findings to further characterise the performance of the deep learning model. First, we report the performance of the deep learning algorithm on an external validation dataset obtained from Stanford University Medical Center, reporting the AUC as well as the sensitivity and specificity at the selected operating point. If there was any discrepancy in sensitivity and specificity, we also planned to present results for these metrics at an operating point selected post hoc to obtain a similar sensitivity (>95%) on the Stanford dataset, to aid in the comparison between primary validation and external validation dataset performance. The model was not retrained or fine-tuned before this assessment. Second, we present the results of the full set of 13 readers, including the non-radiologists and the radiologists who did not interpret the images under diagnostic conditions. Third, we present results of a forced choice experiment, in which equivocal reader responses were treated as definitive for fracture or non-fracture, to further characterise human diagnostic performance, albeit in a manner that does not reflect typical clinical practice. Finally, we report the performance of the deep learning algorithm at clinical prevalence (14%), and the sensitivity and specificity of the algorithm at the selected operating point.

R version 3.62 was used for statistical analysis. p values less than 0·05 were considered to be statistically significant.

Algorithmic audit

We performed a medical algorithmic audit[22,29] by modifying the audit framework described by Raji and colleagues[22] to detect and characterise algorithmic errors, defined as any outputs of the artificial intelligence system that are inaccurate, including those which are inconsistent with the expected performance or which can result in harm if undetected. 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 artificial intelligence system. We then performed a failure mode and effects analysis, and multiple subgroup analyses of the MRMC dataset including a patient-specific subgroup analysis, a task-specific subgroup analysis, and an exploratory error analysis. These subgroup analyses were intended to be descriptive and null hypothesis significance testing was not performed. The methodology and structure of the medical algorithmic audit has been described in detail by Liu and colleagues.[30]

For the task-specific subgroup analysis, the fractures were labelled using a process of schema completion,[21] in which an ontology of clinically relevant fracture subtypes was prospectively defined by a radiologist (LO-R). These subtypes included features regarding the fracture location (eg, subcapital, cervical) and the fracture character (eg, undisplaced, comminuted; appendix p 11). To describe displacement, we used the following system: "subtle" displacement referred to no or minimal cortical step, "mild" displacement referred to up to one cortical width, "moderate" displacement was up to half the bone width, and "severe" displacement was more than half the bone width. We did not distinguish between translation and angulation or tilt, but instead referenced only the most displaced component or region of the fracture. We felt that this

descriptive ontology best described the useful elements of visual variation in the x-rays. Performance was reported in these subgroups and compared against the performance of human readers using the ROC-AUC of the model and the SROC-AUC of the readers. For the exploratory error analysis, a qualified radiologist reviewed all errors the model produced to identify common patterns (failure modes) within the MRMC dataset. This involved visual inspection of the image data, as well as the use of Grad-CAM saliency maps[31] to better characterise model behaviour. The Central Adelaide Local Health Network provided ethical approval for this study (R20171104 HREC/17/RAH/480) and patient consent was waived. The Stanford institutional review board waived the need for approval.

Role of the funding source
There was no funding source for this study.

Results

The numbers of cases and images in the Royal Adelaide Hospital development, primary validation, and the MRMC dataset, and the external validation dataset are shown in figure 1, with dataset characteristics shown in table 1.

In the primary performance comparison (the reader study), the model AUC was 0·994 (95% CI 0·988–0·999), and the AUC of the SROC for the five radiologists was 0·969 (0·960–0·978; figure 2).

A confusion matrix showing the number of false positive and false negative errors is presented in the appendix (p 1); nine false negatives and one false negative were found in the test set of 200 fractures and 200 non-fractures. In a simulated forced choice experiment, in which all definite or equivocal fracture responses from the readers were treated as a positive finding (instead of only the definite responses), reader performance was higher, with an SROC-AUC of 98·5 (95% CI 0·958–1·000), although this approach does not reflect clinical practice (appendix p 2).

Our model achieved an AUC of 0·980 (95% CI 0·931–1·000) on the external validation dataset, which was not significantly different from the results reported on the primary validation dataset (p=0·20). However, the operating point (0·62) did not produce a similar balance of sensitivity and specificity as on the primary validation set when the model was applied to the external validation dataset, producing a sensitivity of 75·0 and a specificity of 100·0 (*vs* sensitivity of 95·5 and specificity of 99·5 in the MRMC dataset). In a 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%.

The sensitivity and specificity of the deep learning model at the preselected operating point (0·62) and the performance of the model at clinical prevalence (14%) are presented in the appendix (p 3), with an unchanged AUC of 0·994, a sensitivity of 94·5%, and a specificity of 99·1%. The full set of audit artefacts are provided in the appendix (pp 6–19), including the full failure mode and effects analysis documents.

The performance of the additional readers in the primary reader study is shown in figure 3; the SROC-AUC for radiologists using non-diagnostic conditions was 0·943 and for non-radiologists was 0·902.

The performance of our artificial intelligence model for demographic subgroups (patient-specific subgroup analysis) is presented in table 2 and for clinically relevant fracture subgroups (task-specific subgroup analysis) is presented in table 3. Patient race and ethnicity information was not available. These analyses show no obviously aberrant model behaviour. Although the performance of the model is slightly lower in the oldest patient cohort (age >80 years), a similar reduction in diagnostic accuracy is

	Development dataset (n=18 178)	Primary validation dataset (n=2449)	MRMC dataset (n=400)	External validation dataset (n=81)
Frontal pelvic x-rays	32 182	2490	400	NA
Unilateral hip images	45 786	4577	400	81
Age, years	69·9 (22·0)	63·7 (25·4)	74·3 (24·0)	63·5 (23·5)
Sex				
Female	9543 (52%)	1178 (48%)	242 (60%)	50 (62%)
Male	8725 (48%)	1271 (52%)	158 (40%)	31 (38%)
Images referred from emergency department	15 127 (47%)	2490 (100%)	400 (100%)	NA
Fracture prevalence	4861 (11%)	356 (14%)	200 (50%)	40 (49%)

Data are n, mean (SD), or n (% of patients). MRMC=multi-reader, multi-case. NA=not applicable.

Table 1: **Dataset characteristics**

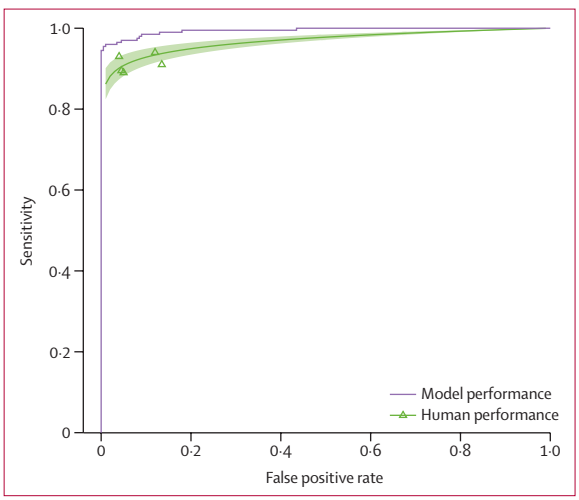

Figure 2: **Primary performance comparison**
Green triangles show the individual performance of the five radiologists. The solid green line shows the average human performance with a summary ROC curve, with the 95% confidence region shown by the shaded area. The purple line shows the ROC curve for the performance of the deep learning model. ROC=receiver operating characteristic.

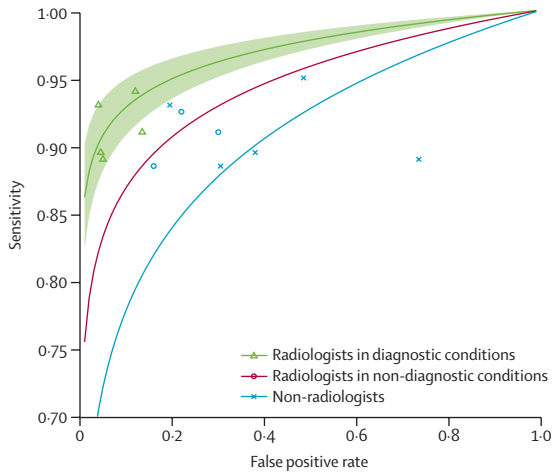

Figure 3: **Additional performance results for other readers**
Symbols show the individual performance of each reader. The summary receiver operating characteristic curves are shown for the primary reader study radiologists in diagnostic conditions (n=5), additional radiologists using non-diagnostic monitors (n=3), and non-radiologists (n=5).

	Cases	AUC	
		Artificial intelligence model	Radiologists
Male	160	0·996	0·979
Female	240	0·994	0·967
Age <40 years	53	0·993	0·970
Age 40–60 years	63	1·000	0·992
Age 61–80 years	59	0·999	0·998
Age >80 years	225	0·988	0·967
Overall performance	400	0·994	0·969

AUC=area under the receiver operating characteristic curve.

Table 2: **Patient-specific subgroup analysis**

	Cases	AUC	
		Artificial intelligence model	Radiologists
Subtle fractures	9	0·964	0·982
Mild displacement	61	0·998	0·969
Moderate displacement	56	1·000	0·990
Severe displacement	74	1·000	0·946
Comminuted fracture	75	1·000	0·971
Subcapital location	66	0·999	0·980
Cervical location	23	0·984	0·982
Pertrochanteric location	105	0·999	0·958
Subtrochanteric location	6	0·970	0·968
Overall performance	400	0·994	0·969

AUC=area under the receiver operating characteristic curve.

Table 3: **Task-specific subgroup analysis**

seen among the radiologists. Furthermore, model performance does not decrease substantially for intracapsular fractures (subcapital and cervical locations), which have distinct clinical implications.

In the exploratory error analysis, other than the subgroups already identified in the patient-specific subgroup analysis and the task-specific subgroup analysis, cases with abnormal bone or joint appearances were noted to be overrepresented among the errors for the model (appendix pp 11–13). Targeted relabelling of the MRMC was done, which revealed six cases with abnormal trabecular patterns due to Paget's disease of the pelvis or femur, or severe femoral head deformities.

A subgroup analysis was performed, and although limited by the modest number of cases involved, there was a large difference in the error rates for the overall MRMC dataset (error rate of 2·5%) and the cases with abnormal bones and joints (error rate of 50·0%; appendix p 15).

No other obvious subgroups were identified during exploratory error analysis. One further surprising error occurred, which was a false negative in a significantly displaced fracture (appendix p 12). The remaining six false negatives (ie, those not already presented on appendix pp 12, 15) are also shown in the appendix (p 5).

On review of Grad-CAM saliency maps,[31] the model was noted to have a tendency to focus on the inner cortex of the neck of the 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 (appendix p 4), even when the model correctly diagnosed the fracture. In the example in the appendix (p 12), the outer cortex is clearly disrupted, but a plausible intact curve along Shenton's line is able to be discerned. This could reflect a failure mode of the model: if displaced fracture elements form a pseudo-Shenton's line, our model might misinterpret this as an intact hip. However, little can be determined from this single error.

Discussion

We report a thorough investigation of a high-performance deep learning algorithm for the detection of proximal femoral fractures from frontal pelvic radiographs. The deep learning model achieved high performance, outperforming radiologists in diagnostic reporting conditions on both the primary metric (AUC 0·994 *vs* 0·969) and by showing both higher sensitivity (95·5% *vs* 94·5% for the best radiologist) and specificity (99·5% *vs* 97·5%) than any doctor tested in the reader study. We also note that the model's performance was higher than that reported by Krogue and colleagues,[33] perhaps due to the smaller dataset and image down-sampling used in their research. To investigate concerns that preclinical artificial intelligence testing can obscure various problems with artificial intelligence models and

lead to a so-called implementation gap,[13] we performed a series of secondary analyses and an algorithmic audit.

In terms of generalisability, our external validation results from a US cohort were informative. Whereas the discriminative performance of the artificial intelligence system (the AUC) appears to be maintained on external validation, the decrease in sensitivity at the prespecified operating point (from 95·5 to 75·0) would make the system clinically unusable in the new environment. Although this shift could be mitigated by the selection of a new operating point, as shown when we found similar sensitivity and specificity in a post-hoc analysis (in which the smaller decrease in specificity reflects the minor reduction in discriminative performance), this would require a localisation process to determine the new operating point in the new environment. To our knowledge, this is the first report of such behaviour in the medical artificial intelligence literature.

Given the tendency of artificial intelligence models to behave in unexpected ways (ie, unlike a human expert would), the inclusion of an algorithmic audit appears to be informative. As stated by Liu and colleagues,[30] the audit approach changes the focus from evaluating the best performance an artificial intelligence system can achieve to identifying the worst mistake it could make. Identifying the types of cases an artificial intelligence model fails on might assist in bridging the current gap between apparent high performance in preclinical testing and challenges in the clinical implementation of an artificial intelligence model.

We note that although our model shows high performance, and does not appear to deviate from human performance in prespecified subgroups, it does still make the occasional inhuman error (eg, misdiagnosing a highly displaced fracture). We also note on saliency mapping that although the model reproduces some recognisable aspects of human practice (eg, it appears to pay attention to Shenton's line), the visualisations nonetheless raise concerns about the regions that are 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. Saliency maps should be interpreted with caution due to known failings of these methods,[34] but these findings together raise the concern that, despite the model performing extremely well at the task of proximal femoral fracture detection when assessed with summary statistics, the model appears to be prone to making unexpected mistakes and can behave unpredictably on cases that humans would consider simple to interpret. These results will hopefully be useful when planning to integrate our model into clinical workflows. Possible strategies to mitigate various issues have been suggested in the algorithmic audit report, such as detailed planning discussions with the relevant clinical teams to consider the effect of the algorithm on care pathways (appendix pp 16–19).

Our study had a number of limitations. First, the deep learning model itself is limited by being unable to act on cases with implanted metalwork (although our system is able to automatically identify these cases and exclude them from analysis). Second, 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 five radiologists reviewed the cases under diagnostic conditions as defined in the local standards of practice. However, the sample size is similar to that in other similar studies[11] and the 95% CIs are not excessively wide. Similarly, the sample size for the external validation is modest but, again, the CIs are reassuring from a clinical perspective. Third, we were unable to access racial identity or ethnicity data for our local population for subgroup testing, and as such were unable to evaluate the stability of the model performance on groups with different racial and ethnic backgrounds.

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

Our study evaluated a high-performance proximal femoral fracture detection deep learning model, which outperforms highly trained clinical specialists in diagnostic conditions, as well as other clinical readers in normal clinical conditions. The performance of the artificial intelligence system was maintained when applied to an external validation sample, and a thorough analysis of the behaviour of the artificial intelligence system shows that it is mostly consistent with that of human experts. We also characterised the occasional aberrant or unexpected behaviour of the artificial intelligence model which could inform future clinical testing protocols. We next intend to test our model in a clinical environment, in the form of an interventional randomised controlled trial.

Contributors

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

Declaration of interests

GC, LJP, and APB acknowledge the support received by the Australian Research Council's Discovery Projects funding scheme (project DP180103232). All other authors declare no competing interests.

Data sharing

The image data used to train and test the artificial intelligence model are not shareable under the current agreement with the data custodian (SA Health). The derived data, including the model and de-identified 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.

Acknowledgments

We thank all of the clinicians who generously donated their time during the reader study.

References

1 Brauer CA, Coca-Perraillon M, Cutler DM, Rosen AB. Incidence and mortality of hip fractures in the United States. *JAMA* 2009; **302:** 1573–79.

2 Kannus P, Parkkari J, Sievänen H, Heinonen A, Vuori I, Järvinen M. Epidemiology of hip fractures. *Bone* 1996; **18** (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; **12:** 366–69.

4 Cannon J, Silvestri S, Munro M. Imaging choices in occult hip fracture. *J Emerg Med* 2009; **37:** 144–52.

5 Pincus D, Ravi B, Wasserstein D, et al. Association between wait time and 30-day mortality in adults undergoing hip fracture surgery. *JAMA* 2017; **318:** 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; **48:** 1155–58.

7 Simunovic N, Devereaux PJ, Bhandari M. Surgery for hip fractures: does surgical delay affect outcomes? *Indian J Orthop* 2011; **45:** 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; **26:** 1197–99, discussion 1199.

9 Nagendran M, Chen Y, Lovejoy CA, et al. Artificial intelligence versus clinicians: systematic review of design, reporting standards, and claims of deep learning studies. *BMJ* 2020; **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; published online Nov 17. https://arxiv.org/abs/1711.06504 (preprint).

11 Liu X, Faes L, Kale AU, et al. A comparison of deep learning performance against health-care professionals in detecting diseases from medical imaging: a systematic review and meta-analysis. *Lancet Digit Health* 2019; **1:** e271–97.

12 Kohli A, Jha S. Why CAD failed in mammography. *J Am Coll Radiol* 2018; **15:** 535–37.

13 Seneviratne MG, Shah NH, Chu L. Bridging the implementation gap of machine learning in healthcare. *BMJ Innov* 2020; **6:** 45–47.

14 Lehman CD, Wellman RD, Buist DSM, Kerlikowske K, Tosteson AN, Miglioretti DL. Diagnostic accuracy of digital screening mammography with and without computer-aided detection. *JAMA Intern Med* 2015; **175:** 1828–37.

15 Lyell D, Coiera E. Automation bias and verification complexity: a systematic review. *J Am Med Inform Assoc* 2017; **24:** 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; **15:** e1002683.

17 Nam JG, Park S, Hwang EJ, et al. Development and validation of deep learning-based automatic detection algorithm for malignant pulmonary nodules on chest radiographs. *Radiology* 2019; **290:** 218–28.

18 McKinney SM, Sieniek M, Godbole V, et al. International evaluation of an AI system for breast cancer screening. *Nature* 2020; **577:** 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. In: Petersen J, Estépar RSJ, Schmidt-Richberg, et al (eds). Thoracic image analysis. Cham: Springer, 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; published online Sept 23. https://arxiv.org/abs/2009.11060 (preprint).

21 Oakden-Rayner L, Dunnmon J, Carneiro G, Ré C. Hidden stratification causes clinically meaningful failures in machine learning for medical imaging. *Proc ACM Conf Health Inference Learn* 2020; **2020:** 151–59.

22 Raji ID, Smart A, White RN, 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: Association for Computing Machinery, 2020: 33–44.

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. Piscataway, NJ: Institute of Electrical and Electronics Engineers, 2017: 4700–08.

24 Varma M, Lu M, Gardner R, et al. Automated abnormality detection in lower extremity radiographs using deep learning. *Nat Mach Intell* 2019; **1:** 578–83.

25 The Royal Australian and New Zealand College of Radiologists. Standards of practice for diagnostic and interventional radiology, version 11.2. 2020. https://www.ranzcr.com/documents/510-ranzcr-standards-of-practice-for-diagnostic-and-interventional-radiology/file (accessed Sept 3, 2020).

26 Irwig L, Tosteson AN, Gatsonis C, et al. Guidelines for meta-analyses evaluating diagnostic tests. *Ann Intern Med* 1994; **120:** 667–76.

27 Gatsonis C, Paliwal P. Meta-analysis of diagnostic and screening test accuracy evaluations: methodologic primer. *AJR Am J Roentgenol* 2006; **187:** 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; **44:** 837–45.

29 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; **27:** 132–35.

30 Liu X, Glocker B, McCradden M, Ghassemi M, Denniston AK, Oakden-Rayner L. The medical algorithmic audit. *Lancet Digit Health* 2022; published online April 5. https://doi.org/10.1016/S2589-7500(22)00003-6.

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. Piscataway, NJ: Institute of Electrical and Electronics Engineers, 2017: 618–26.

32 Jones DHA. Shenton's line. *J Bone Joint Surg Br* 2010; **92:** 1312–15.

33 Krogue JD, Cheng KV, Hwang KM, et al. Automatic hip fracture identification and functional subclassification with deep learning. *Radiol Artif Intell* 2020; **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, eds. Advances in neural information processing systems 31 (NeurIPS 2018). Red Hook, NY: Curran Associates, 2018: 9505–15.