



Published in final edited form as:

Pediatr Radiol. 2025 May ; 55(6): 1257–1269. doi:10.1007/s00247-025-06223-4.

Using deep learning for estimation of time-since-injury in pediatric accidental fractures

Farah W. Brink^{1,2}, Brent Adler³, Sven Bambach³, Charmaine B. Lo^{1,3}, Steven Rust³, Christopher W. Bartlett^{1,2}, Logan Bradshaw¹, M. Katherine Henry⁴, Diana Messer⁵

¹Nationwide Children's Hospital, 700 Children's Drive, Columbus, OH 43205, USA

²The Ohio State University Wexner Medical Center, Columbus, USA

³Abigail Wexner Research Institute at Nationwide Children's Hospital, Columbus, USA

⁴Clinical Futures, Division of General Pediatrics, and Department of Radiology, Children's Hospital of Philadelphia, Philadelphia, USA

⁵University of Tennessee Health Science Center, Memphis, USA

Abstract

Background: Estimating time-since-injury of healing fractures is imprecise, encompassing excessively wide timeframes. Most injured children are evaluated at non-children's hospitals, yet pediatric radiologists can disagree with up to one in six skeletal imaging interpretations from referring community hospitals. There is a need to improve image interpretation by considering additional methods for fracture dating.

Objective: To train and validate deep learning models to correctly estimate the age of pediatric accidental long bone fractures.

Materials and methods: This secondary data analysis used radiographic images of accidental long bone fractures in children <6 years at the time of injury seen at a large Midwestern children's hospital between 2000–2016. We built deep learning models both to classify fracture images into different age groups and to directly estimate fracture age (time-since-injury). We used cross-validation to evaluate model performance across various metrics, including confusion matrices, sensitivity/specificity, and activation maps for age classification, and mean absolute error (MAE) and root mean squared error (RMSE) for age estimation.

Farah W. Brink farah.brink@nationwidechildrens.org.

Author contributions F.B. conceptualized and designed the study, drafted the initial manuscript, and reviewed and revised the manuscript. B.A. reviewed and annotated images and reviewed and revised the manuscript for important intellectual content. S.B. carried out data management and all analyses, designed the study, assisted with drafting the initial manuscript, and reviewed and revised the manuscript. C.L. assisted with drafting the initial manuscript and critically reviewed and revised the manuscript for important intellectual content. S.R. designed the study and reviewed and revised the manuscript for important intellectual content. C.B. and M.H. reviewed study design and critically reviewed the manuscript for important intellectual content. L.B. assisted with data management, assisted with drafting the initial manuscript, and reviewed and revised the manuscript. All authors approved the final manuscript as submitted and agree to be accountable for all aspects of the work.

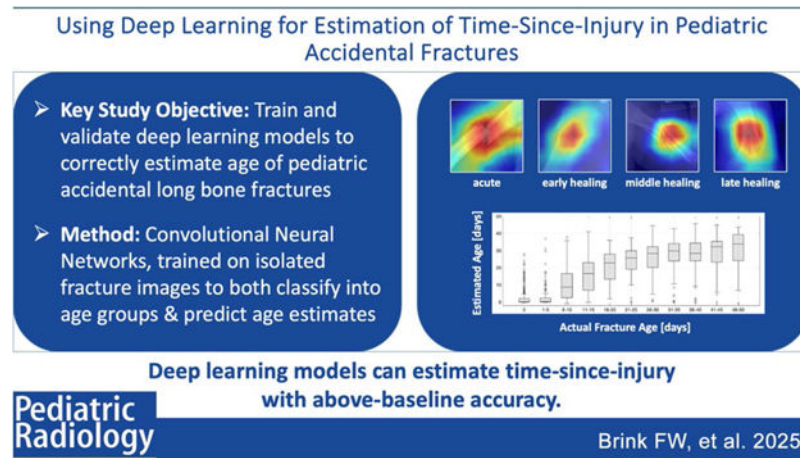
Declarations

Competing interests Drs. Farah Brink and M. Katherine Henry's institutions have been paid in cases with concern for child maltreatment for which they were subpoenaed to provide expert testimony.

Results: Our study cohort contained 2,328 radiographs from 399 patients. Overall, our models performed above baselines for fracture age classification and estimation, both when trained/validated across all bones and on specific bone types. The best model was able to estimate fracture age for any long bone with a MAE of 6.2 days and with 68% of estimates falling within 7 days of the correct fracture age.

Conclusion: Our study successfully demonstrated that, for radiographic dating of accidental fractures of long bones, deep learning models can estimate time-since-injury with above-baseline accuracy.

Graphical Abstract



Keywords

Artificial intelligence; Bone; Children; Dating; Deep learning; Fractures; Radiographs

Introduction

Estimating time-since-injury of healing fractures is imprecise and can encompass excessively wide timeframes [1]. The radiographic depiction of the features of healing including subperiosteal new bone, callus, and remodeling appear over time and represent multiple physiologic processes, including the mineralization and ingrowth of osteocytes in a fracture hematoma and the induction of new osteocytes from the damaged periosteum. These processes can project on top of one another and can be difficult to distinguish in many instances. Additionally, although some features of healing can occur earlier than others, these features constitute continuous processes that are expected to begin during wide timeframes of days to weeks, making fracture age estimation for radiologists a challenging task [2]. Several proposed timelines to guide fracture dating have been reported, relying mostly upon accidental and birth injuries to determine radiographic features of healing [1, 3–9]. Pediatric radiologists generally use knowledge and experience to estimate healing fracture age. As most injured children are evaluated at non-children's hospitals, and pediatric radiologists can disagree with up to one in six skeletal imaging interpretations

from referring community hospitals, there is an urgent need for efforts to improve image interpretation [10, 11].

Machine learning is a branch of artificial intelligence (AI) with extraordinary potential to analyze patterns in data [12, 13]. Deep learning, a type of machine learning based on multi-layered artificial neural networks, can learn to recognize patterns in images and use them to draw high-level conclusions [13]. Deep learning models have been applied to a variety of radiologic tasks in the care of adults, though it has yet to be widely investigated for the purpose of accurately estimating time-since-injury of fractures. Tsai et al. recently evaluated the performance of a deep learning model for dating birth-related clavicle fractures and found that 83.7% of age estimates were accurate within 7 days of the actual age of the fracture (e.g., ground truth) [14]. However, there were several limitations to this study including small numbers of fractures in one bone type. The authors additionally manually aligned the images for both the training and validation cohorts which may not mimic a real-world scenario. Moreover, timelines derived from healing of birth trauma may not be accurate when extrapolated to fracture healing in older infants and children, as healing trajectories likely differ [9].

At present, estimating fracture age is challenging and existing methods are imprecise, inconsistent, and unstandardized [1]. The development of models to age healing fractures within long bones has the potential to serve as a virtual consultant to radiologists particularly in a community setting. The objective of our study was to assess the feasibility of developing deep learning models that can classify accidental fractures of the clavicle and long bones of the upper and lower extremities in children <6 years into different fracture age categories, and to also estimate time-since-injury (fracture age) directly.

Methods

We utilized a retrospective dataset of long-bone fracture radiographs with known fracture ages and systematically trained deep learning models to estimate the fracture age based on only the fracture image. We followed the CLAIM checklist for AI in Medical Imaging [15]. This single-institution retrospective study was approved by our local institutional review board.

Study design/population

We conducted a secondary data analysis using radiographic images (anterior-posterior and lateral views) that were collected as part of a previous study of accidental fractures of long bones (clavicle, humerus, radius, ulna, femur, tibia, fibula) in children <6 years at the time of injury and seen at a large Midwestern children's hospital between 2000–2016 [16]. This convenience sample of children was initially collected for a study examining the radiographic characteristics of healing long bone fractures in accidental injury in young children [16]. Radiographic images were identified via the hospital's Trauma Registry, electronic health record, the child protection team administrative dataset and the local picture archiving and communication system (PACS). After assessment on the PACS, images were saved as PNG files. The ground truth of the fracture age was determined by chart review. Patients were excluded if there was presence of any bone health co-

morbidity and presence of classic metaphyseal lesions, if internal fixation of the fracture was performed, if no follow-up films were found within the PACS, or if physical abuse was suspected by a child abuse pediatrician at the time of evaluation or on review of the medical record. Specifically, any cases of possible physical abuse were excluded due to concerns for omissions by caregivers in the history which could lead to ground truth uncertainty.

Image preparation

All images were reviewed by three of the authors with expertise in pediatric radiology, child abuse pediatrics, and forensic anthropology. Any poor-quality images (e.g., fractures completely obscured by casting material) were excluded. After these exclusions, we included images we considered to be ideal and images that were not ideal because the fracture was partially overlapped by the casting material or an adjacent bone (e.g., the fibula partially obscuring the tibia). We made the decision to train the machine on all the images and then also only on the ideal images to study the effect of including/excluding such fractures during model training. Figure 1 describes the full image inclusion/exclusion flowchart resulting in different analytic samples. Fracture locations were manually identified and annotated with a tight rectangular region of interest (ROI) bounding box. After fracture locations were annotated, a pediatric radiologist re-reviewed all images to ensure accuracy of fracture identification. For neural network processing, this ROI box was padded by increasing its width and height by 50% each to provide the neural network models with some local context around the fracture. Only the padded ROI regions were provided to the models during training and validation. Bounding boxes were clear of any extraneous information such as writing or markers. Figure 2 shows examples of different radiographs and fracture annotations. Note that some radiographs contained two fractures (either two views of the same fracture or two fractures affecting two bones, e.g., tibia and fibula), in which case two ROI regions were annotated (Fig. 2).

Deep learning models

For our deep learning models, we used the well-established convolutional neural network (CNN) architecture ResNet-50, pretrained on ImageNet, and adjusted the network output to suit our two predictive tasks [17, 18]. The first task was to classify fracture age as either acute (0–9 days), early healing (10–20 days), middle healing (21–42 days), or late healing (>42 days). These age groups were determined based on conservative ranges that correspond to timing of healing features as described in the existing literature [1, 3–9, 19]. For this task, we used four output neurons and a cross-entropy loss function, with per-class weights such that each class has equal contribution to the loss, thus preventing the model from simply learning the class prevalence. The second task was to directly estimate fracture age (in days), for which we used a single output neuron and either mean squared error (MSE) or mean absolute error (MAE) as the loss function.

Model implementation and experiments

All models and experiments were implemented in Python using the PyTorch (version 1.10) deep learning library [20]. Model training was done on a high-performance computing cluster using either an NVIDIA A100 or V100 GPU, depending on availability. Optimization was done using the Adam optimizer with a learning rate of 1e-4 and a batch

size of 128 and ended automatically once the validation loss had not improved for three epochs, restoring the model parameters with the smallest validation loss up until then [21]. Data augmentation was applied on the fly and included randomly cropping images to 224×224 pixels before rescaling them to the 256×256 pixel input required for ResNet-50.

We conducted multiple experiments by fitting separate models either to the full analytic sample (all bone types) or to an ideal sample without overlap or casting obscuring the fracture (all bone types), as well as fitting individual models to different groups of bones (clavicle, humerus, radius/ulna, femur, tibia/fibula) (Figs. 1 and 2).

Model evaluation

Given the limited amount of data, we evaluated each experiment using 5-fold cross-validation, which we then repeated five times, reporting mean performance and 95% confidence intervals (CIs) across repetitions. The motivation for the latter was to capture variability caused by the inherent randomness of training deep neural network models. Because one patient could have multiple radiographs of the same fracture at different stages of healing, we performed the cross-validation at the patient-level, ensuring that models are never trained and validated on radiographs from the same patient. For the fracture age classification task, interpreting and comparing overall classification accuracies would be challenging because baseline accuracies depend on the class prevalence, which differs across experiments. We instead report class-based sensitivities and specificities, as well as confusion matrices. We also present selected image heatmaps using gradient-based Class Activation Mapping (grad-CAM), which is a common technique to localize regions in an image that had the most influence on the classification output of a CNN model [22]. For the fracture age estimation task, we report the mean absolute error (MAE) and the root-mean-square error (RMSE), as well as the percentage of age estimates that were within 7 days of ground truth.

Results

Bone fracture cohort

Table 1 and Fig. 1 summarize our bone fracture cohort for both the full analytic sample, including the number of ROI images obscured by casting or obscured by other bone, and the ideal sample. A total of 399 patients with a total of 2,328 radiographs were included; the mean age of patients on the first day of imaging was 2.6 years with a standard deviation of 1.6 years. Eighteen percent of patients ($n=70$) were less than 1 year old, 23% ($n=93$) were 1–2 years old, 16% ($n=63$) were 2–3 years old, 18% ($n=70$) were 3–4 years old, 14% ($n=54$) were 4–5 years old, and 12% ($n=49$) were 5–6 years old. The greatest number of fracture ROI images was for the humerus (568 images from 90 patients) and the least number for the fibula (159 images from 31 patients). On average, patients had radiographs taken at 3.0 different stages of healing. Additionally, some radiographs contained two views of the same fracture or fractures affecting two bones (e.g., radius and ulna), both resulting in multiple fracture ROI images, adding to a total of 6.2 ROI images per patient on average. Figure 3 contains the distribution of fracture age, showing a peak of images taken on the day of injury

(0 days), with a somewhat uniform distribution thereafter through a fracture age of 45–50 days. This distribution shape was consistent among all types of bones.

Fracture age classification

Figure 4 shows the confusion matrices when classifying fractures from all bone types as either acute (0–9 days), early healing (10–20 days), middle healing (21–42 days), or late healing (>42 days). For the analytic sample, the plurality of images within each class was assigned correctly and confusion was more often between adjacent classes (e.g., middle versus late healing) as opposed to non-adjacent classes such as acute versus late healing. When training and evaluating on the ideal sample only, the matrix looks similar except for increased confusion within the early healing class of fractures.

Figure 5 shows representative grad-CAM heatmaps of fractures that were correctly classified based on the model trained on the ideal sample (all bones). The heatmaps suggest that the deep learning model did primarily focus on the area of the fracture to make its classification. In addition, Supplementary Fig. 1 shows heatmaps of incorrectly classified fractures and Supplementary Fig. 2 shows heatmaps for images from the analytic sample that contain casting over the fracture.

Table 2 breaks down the per-class sensitivity and specificity when training/evaluating models on the full analytic sample (i.e., all bone types), the ideal sample, and different groups of bones. Figure 6 plots these values and confidence regions employing receiver operating characteristic (ROC) curve axes to visually summarize results (confidence regions above the dashed line imply above-chance model performance). When training on all bones (analytic sample and ideal sample), model performance was above chance for all four fracture age classes. Most of the models trained on specific bone groups also performed above chance, although with larger confidence intervals due to the smaller numbers of images, especially when reducing the dataset to the ideal sample. For the early healing stage of the radius/ulna, tibia/fibula, and femur, the numbers of ideal images were so small that no meaningful analysis was possible.

Fracture age estimation

Table 3 summarizes the results of fracture age estimation for the analytic sample, ideal sample, and different bone groups, both for models trained on MAE and for models trained on MSE. Notably, for this task, we excluded outlier images with a fracture age of 50 days or more (Fig. 1). The estimated age errors were better than educated baselines (i.e., guessing the median/mean age of each dataset) in all experiments except for the group of ideal femur fractures, which was also the smallest dataset ($n=154$ images). Overall, the lowest error was achieved when training/validating on the ideal sample (all bones), with a MAE of 6.17 days and a RMSE error of 9.42 days, corresponding to 68% of bone fracture estimates to be within 7 days of the true age. Note that while this error was somewhat affected by the high proportion of acute 0-day-old fractures, the same model still performed better than the baseline after all 0-day images were removed (MAE=8.23, baseline=12.62). Figure 7 summarizes fracture age estimation results for the ideal sample. The estimated distribution of fracture age is similar in shape to the actual age distribution. The scatterplot of actual

versus estimated fracture age demonstrates a strong correlation (Pearson $r=0.80$); however, there are instances where the estimates deviate considerably from the actual values (e.g., a 44-day-old fracture being estimated as acute) (Fig. 7). Boxplots of estimated fracture age grouped by actual age show that, while the mean estimated age increases with actual age, the range of estimates is still broad (e.g. 5–45 days for fractures that are between 26–30 days old) (Fig. 7).

Bone-specific models versus general models

Supplementary Tables 1 and 2 include age classification sensitivity/specificity (Supplementary Table 1) and age estimation errors (Supplementary Table 2) of different models trained on all bone types, categorized by bone group. This allows a direct comparison to the models that were trained on separate bone groups. For age estimation, models trained on all bones had lower error rates than those trained on a specific bone group in almost all cases, with the biggest improvements for bones with smaller amounts of data. For fracture age classification, the advantage of the model trained on all bones was present but less consistent.

Discussion

In this study, we used deep learning models based on pretrained CNNs to explore two separate approaches to present an estimated fracture age of long bones: grouping into different age classes and estimating fracture age directly. With over 2,500 fracture images from 399 patients, we trained and tested our CNN models on a cohort substantially larger than in prior studies exploring fracture age estimation [3–9, 14]. Our dataset is also the first to include accidental fractures of multiple bones in children <6 years instead of only focusing on birth trauma in one bone. As the timeframes for healing of various fractures in older children are likely different than in newborns, this key step in training a model is necessary when considering the implications of inappropriately extrapolating healing patterns from one patient age/bone type group to another [9].

We systematically investigated and successfully demonstrated that CNN-based deep learning models can estimate the time-since-injury of accidental fractures of long bones based on radiographs with above-baseline accuracy. Overall, our results indicate that these models produced above-baseline estimates of both fracture age groups (e.g., sensitivity of 65% and specificity of 87% for identifying acute fractures) and numerical fracture age (MAE of 6.2 days compared to a 13.9-day baseline). While deep learning models are often considered black boxes, using grad-CAM maps, we provided some evidence that our models did rely on fracture characteristics to produce their estimates. We also showed that training bone-specific models (with potentially less training data per bone) may not be needed as models trained with all long bones performed as well or better compared to bone-specific models when tested on specific bone types. This may have implications when considering the type and number of images needed for training models in future work.

While we demonstrated that deep learning models can infer time-since-injury of fracture images better than chance and robustly across different bone types, it is important to point out that some model estimates can still be very inaccurate. For example, the classification

model trained on all bone types of our analytic sample still classified 9.8% of acute fractures (<10 days) as late healing (>42 days), and 12.5% of late healing fractures as acute (Fig. 4). When estimating fracture age directly using our ideal sample, 15% ($n=64$) of 0-day-old fractures are estimated to be between 5 days and 28 days old (Fig. 7). Given the complex nature of deep learning models, it is not a simple task to determine why some fracture ages were estimated incorrectly. By training separate models with and without images that contained a cast, we found that models that were trained without casts (ideal sample) were less accurate at identifying early healing fractures. This could be because early healing fracture images were most likely to contain a cast (54% of early healing images had casting versus 21%, 15%, and 6% for acute, middle, and late respectively), and therefore, the early healing class had the most drastic reduction of available training images in the ideal sample. Another explanation could be that the model trained on the analytic sample could have learned that the presence of casting correlates with early healing. Supplementary Figure 2 supports this explanation by showing grad-CAM heatmaps of images containing casting, suggesting the model was less focused on the fracture location and could have been focused on the pattern of the cast instead. Because our models were only exposed to the fracture region of interest and we ensured that those regions were free of any markers or annotations, we did not identify any other obvious visual bias that could have explained incorrect model estimates.

With future implementation in mind, we considered different ways a model's age estimates might be presented to a radiologist in a real-time setting. Rather than providing a single number as the age estimate, providing a probability distribution across four age categories that are similar to current reporting conventions may increase trust in the model and make the radiologist feel more confident in reporting their impression of the estimated fracture age. Moreover, to our knowledge, using deep learning to classify healing phases of bone fractures has not yet been previously published. We believe that future work should further explore accessible ways to present model estimates to providers that maximize clinical utility and facilitate incorporation of the model algorithms into clinical workflows.

For the models that estimated age directly, we used standard loss functions (MAE and MSE), although other loss functions have been suggested specifically for age estimation tasks (e.g., estimating the age of a person given a photo of the face) [23]. We experimented with CORAL loss, similar to Tsai et al., but found that its age estimates were more biased toward extremes, resulting in overall higher errors [14, 24]. We do find it motivating that even standard regression losses resulted in model performance far above chance, leaving the door open for future improvements not only based on more training data, but also based on custom loss functions.

Today, radiologists use their clinical experience and understanding of the few published papers on the radiographic depiction of fracture healing in young children for age estimation. Although adequate for guiding fracture treatment, these methods are insufficient when evaluating for physical abuse where a more accurate determination of fracture age has the potential to provide objective evidence for or against the diagnosis of abuse. Accurate dating of healing fractures can be crucial to investigative agencies as they consider child protection. However, the pattern of healing in abusive fractures may be different

from accidental trauma where the treatment is likely rendered immediately. Given national shortages in pediatric radiologists, additional methods for aiding fracture dating should be considered, particularly in non-pediatric centers, to enhance the ability to identify and protect physically abused victims from further violence.

By using images of accidental fractures for which the time of injury is known, we have established the premise for the development of a computerized tool that could ultimately be used as an aid to radiologists in dating abusive fractures accurately. Our work additionally sets the stage for dating other fractures besides long bones and with higher specificity for physical abuse (e.g., rib fractures and classic metaphyseal lesions). To the best of our knowledge, our study is the first to use artificial intelligence and deep learning to date accidental fractures of long bones in children <6 years. Our results suggest that deep learning models have the potential to be successfully trained and validated in a large population of children and deployed at scale in the form of a computer-aided tool designed to assist providers evaluating potentially abused children with healing bone fractures.

There are limitations to our study. We chose to only train and test models using the region of interest around the fracture rather than full radiographs in an effort to limit potential biases or distractions present in the full image. However, this choice limits practical applications as a similar preprocessing of the radiograph would be necessary. While we were careful to train and validate images at the patient-level, we did not explicitly account for the fact that multiple fracture images from the same patient are not independent, which at worst could lead to overconfident results. While our combined dataset is, to the best of our knowledge, the largest pediatric fracture dataset analyzed with deep learning, the amount of data for specific bones in specific healing stages is still very small and sometimes does not allow us to reach meaningful conclusions (e.g., there were only nine ideal images of a femur fracture during early healing). While we implemented random cropping as a form of data augmentation, there are more data augmentation/synthetization techniques that can be explored in future work. Limited training data is a known challenge for deep learning research in pediatric imaging. Our heterogeneous cohort is no exception to this phenomenon, especially considering the large age range of subjects within our dataset and the physiological differences in bones between an infant and a 5-year-old. The models in this study were built under the assumption that healing would be the same across all ages; yet we acknowledge healing physiology of an infant would differ from that of a 5-year-old. However, due to the small number of image counts in our study once classified by age, this analysis was not feasible and this difference in healing physiology should be considered in future work. Additionally, we were limited by the previously collected dataset only containing PNG files. Future work should focus on the development of large and diverse pediatric imaging datasets with raw high-resolution digital images (e.g., DICOM standard) that provide higher resolution and pixel bit depth.

Conclusion

Our deep learning study successfully demonstrated that models for radiographic dating of accidental fractures of long bones can estimate time-since-injury with above-baseline accuracy. Future multisite collaborations to develop large imaging datasets would allow

for further training and external validation of a fracture age estimation model. This could establish the framework for a computer-aided diagnosis system deployed at scale and designed to assist with dating of healing bone fractures in potentially abused children.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Funding

This research was supported by the authors' primary institution through a grant from the Centers for Disease Control and Prevention (Grant No. 1R49CE003074-01) and by an intramural grant through the authors' primary institution. Dr. Henry's time was supported under grant number (K08HS028847) from the Agency for Healthcare Research and Quality (AHRQ), U.S. Department of Health and Human Services (HHS). The authors are solely responsible for this document's contents, findings, and conclusions, which do not necessarily represent the views of AHRQ, the Centers for Disease Control and Prevention, or the authors' primary institution. Readers should not interpret any statement in this report as an official position of AHRQ or of HHS. None of the authors has any affiliation or financial involvement that conflicts with the material presented in this report unless otherwise reported.

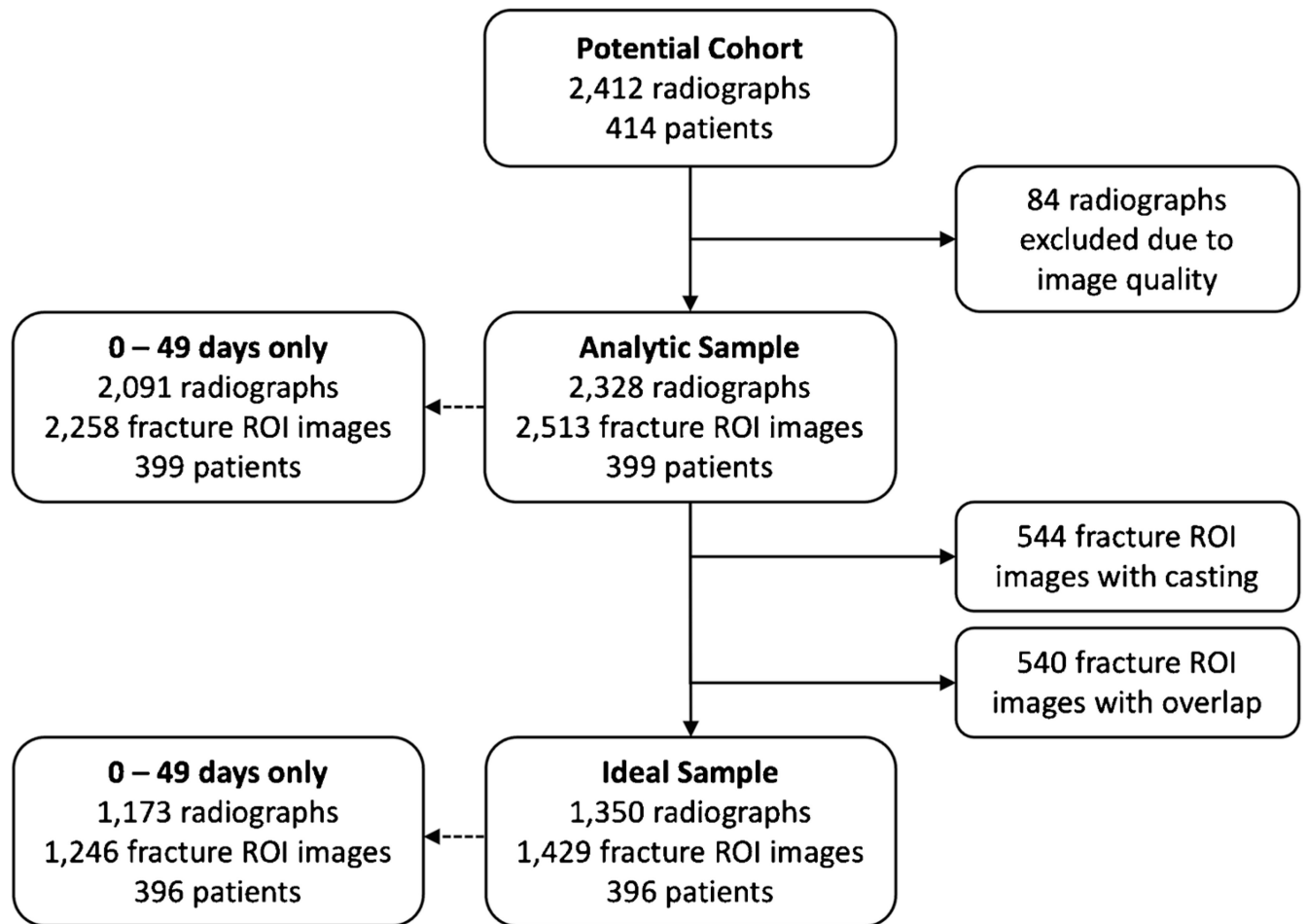
Data availability

The data that support the findings of this study are not openly available due to reasons of sensitivity and are available from the corresponding author upon reasonable request. Data are located in controlled access data storage at Nationwide Children's Hospital.

References

1. Messer DL, Adler BH, Brink FW et al. (2020) Radiographic timelines for pediatric healing fractures: a systematic review. *Pediatr Radiol* 50(8):1041–1048 [PubMed: 32157365]
2. Kleinman PK (Ed.) (2015) Diagnostic imaging of child abuse (3rd ed.). Cambridge University Press. Chapter 6: Dating fractures
3. Yeo LI, Reed MH (1994) Staging of healing of femoral fractures in children. *Canadian Assoc Radiol J J Assoc Canadienne Des Radiol* 45(1):16–19
4. Islam O, Soboleski D, Symons S et al. (2000) Development and duration of radiographic signs of bone healing in children. *AJR Am J Roentgenol* 175(1):75–78 [PubMed: 10882250]
5. Malone CA, Sauer NJ, Fenton TW (2011) A radiographic assessment of pediatric fracture healing and time since injury. *J Forensic Sci* 56(5):1123–1130 [PubMed: 21644992]
6. Halliday KE, Broderick NJ, Somers JM, Hawkes R (2011) Dating fractures in infants. *Clin Radiol* 66(11):1049–1054 [PubMed: 21763645]
7. Prosser I, Lawson Z, Evans A et al. (2012) A timetable for the radiologic features of fracture healing in young children. *AJR Am J Roentgenol* 198(5):1014–1020 [PubMed: 22528890]
8. Warner C, Maguire S, Trefan L et al. (2017) A study of radiological features of healing in long bone fractures among infants less than a year. *Skeletal Radiol* 46(3):333–341 [PubMed: 28070625]
9. Crompton S, Messina F, Klafkowski G et al. (2021) Validating scoring systems for fracture healing in infants and young children: pilot study. *Pediatr Radiol* 51:1682–1689 [PubMed: 33847785]
10. Karmazyn B, Wanner MR, Marine MB et al. (2019) The added value of a second read by pediatric radiologists for outside skeletal surveys. *Pediatr Radiol* 49(2):203–209 [PubMed: 30367201]
11. Densmore JC, Lim HJ, Oldham KT, Guice KS (2006) Outcomes and delivery of care in pediatric injury. *J Pediatr Surg* 41(1):92–98
12. LeCun Y, Bengio Y, Hinton G (2015) Deep learning. *Nature* 521(7553):436–444 [PubMed: 26017442]

13. Zhou SK, Greenspan H, Davatzikos C et al. (2021) A review of deep learning in medical imaging: imaging traits, technology trends, case studies with progress highlights, and future promises. *Proc IEEE Inst Electr Electr Eng* 109(5):820–838
14. Tsai A, Grant PE, Warfield SK et al. (2022) Deep learning of birth-related infant clavicle fractures: a potential virtual consultant for fracture dating. *Pediatr Radiol* 52(11):2206–2214 [PubMed: 35578043]
15. Tejani AS, Klontzas ME, Gatti AA et al. (2024) Checklist for Artificial Intelligence in Medical Imaging (CLAIM): 2024 update. *Radiol Artif Intell* 6(4):e240300
16. Messer DL (2019) Variables influencing time since injury of pediatric healing fractures; radiographic assessment and implications for child physical abuse [Doctoral dissertation, Ohio State University]. OhioLINK Electronic Theses and Dissertations Center. https://etd.ohiolink.edu/acprod/odb_etd/etd/r/1501/10?clear=10&p10_accession_num=osu1574408203228578
17. He K, Zhang X, Ren S, Sun J (2016) Deep residual learning for image recognition. In: *Proceedings of the IEEE conference on computer vision and pattern recognition* pp 770–778
18. Kornblith S, Shlens J, Le QV (2019) Do better imagenet models transfer better?. In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* pp 2661–2671
19. Paddock M, Choudhary AK, Jeanes A et al. (2023) Controversial aspects of imaging in child abuse: a second roundtable discussion from the ESPR child abuse taskforce. *Pediatr Radiol* 53(4):739–751 [PubMed: 36879046]
20. Paszke A, Gross S, Massa F et al. (2019) PyTorch: an imperative style, high-performance deep learning library. *ArXiv*, abs/1912.01703
21. Kingma DP, Ba J (2014) Adam: a method for stochastic optimization. *ArXiv preprint*. arXiv:1412.6980
22. Selvaraju RR, Cogswell M, Das A et al. (2017) Grad-cam: visual explanations from deep networks via gradient-based localization. In: *Proceedings of the IEEE international conference on computer vision* pp 618–626
23. Niu Z, Zhou M, Wang L et al. (2016) Ordinal regression with multiple output CNN for age estimation. 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). IEEE pp 4920–4928
24. Cao W, Mirjalili V, Raschka S (2020) Rank consistent ordinal regression for neural networks with application to age estimation. *Pattern Recog Lett* 140:325–331

**Fig. 1.**

Flowchart illustrating the two samples of images analyzed. We use both the analytic sample and the ideal sample (excludes images where the fracture is obscured by a cast or other bones) in our analysis. A small number of radiographs contained either two views of the same fracture or two fractures that affected two bones, resulting in more fracture ROI images than radiographs. Specifically, in our analytic sample, 86 radiographs contained two views of the same fracture, 77 radiographs contained fractures affecting the radius and ulna, and 22 radiographs contained fractures affecting both the tibia and fibula. In some experiments (i.e., direct fracture age estimation), we limited the data to fractures that were less than 50 days old

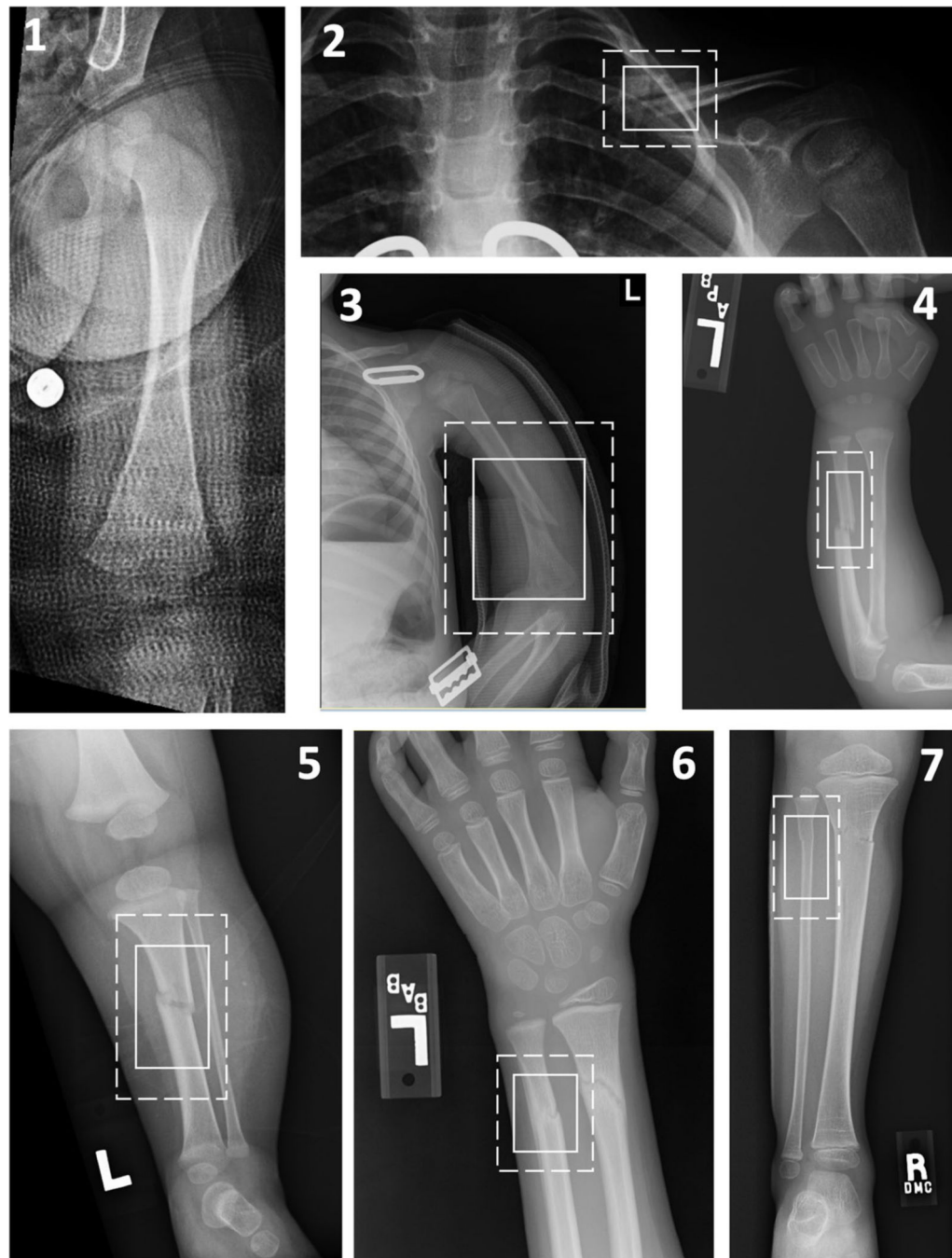


Fig. 2.

Example radiographs of different fractures. (1) Femur, excluded due to image quality. (2) Clavicle with bone overlap. (3) Humerus with cast. (4-7) Ideal examples of radius, tibia, ulna, and fibula, respectively. *Solid boxes* are the annotated fracture regions of interest. *Dashed boxes* are the padded regions provided to the neural network models. Note that if a radiograph contained fractures of two bones (e.g., radius and ulna in example 6), two *boxes* were created and treated as separate images thereafter. Only one *box* is shown here for ease of visualization

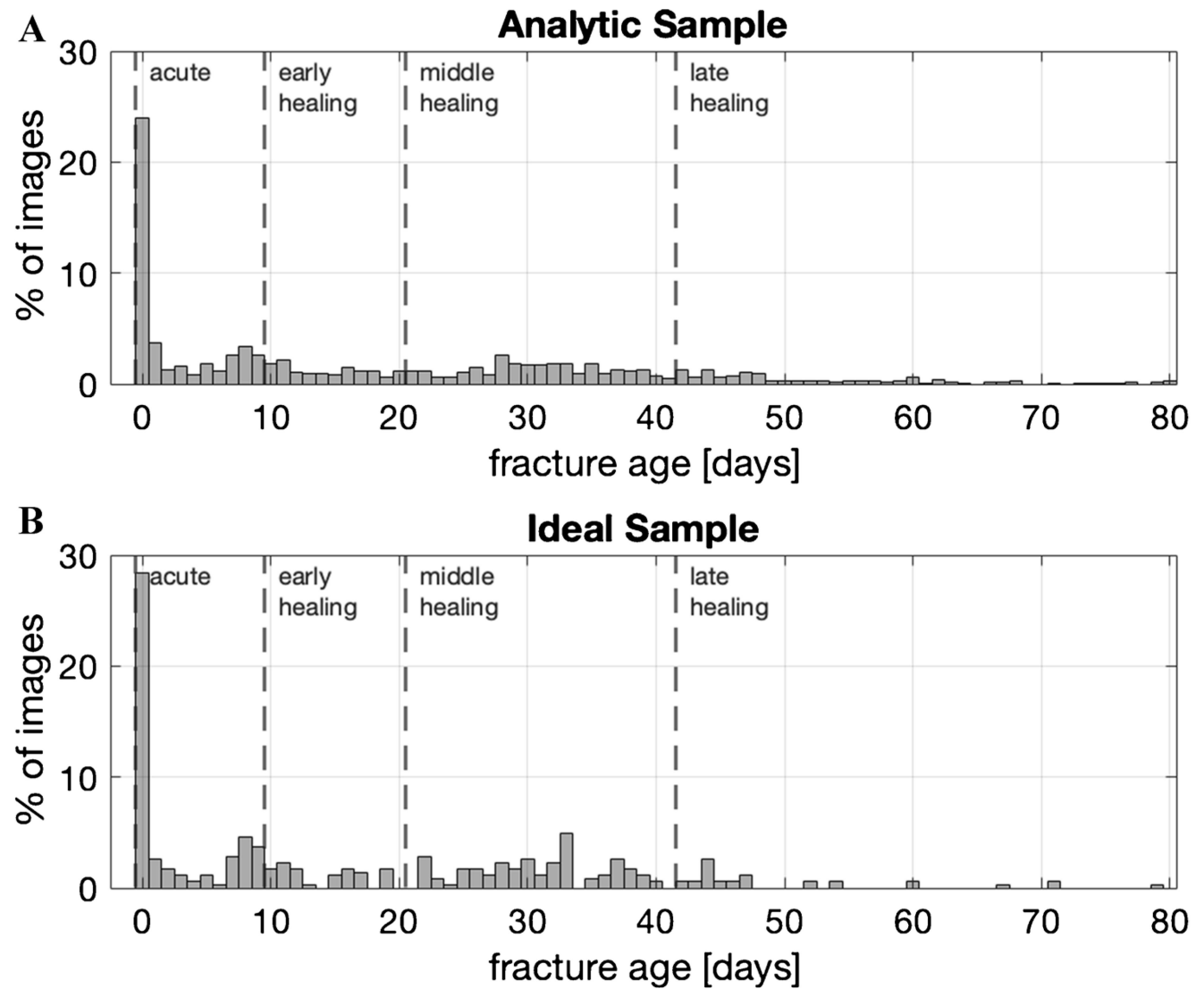


Fig. 3. Distribution of fracture age in our analytic sample (2,513 images, 449 patients) (**a**) and ideal sample (1,429 images 430 patients) (**b**). Note that outliers >80 days old (4.9% and 0.3% of images for analytic and ideal samples, respectively, with a maximum of 451 and 127 days) are not shown here for better readability

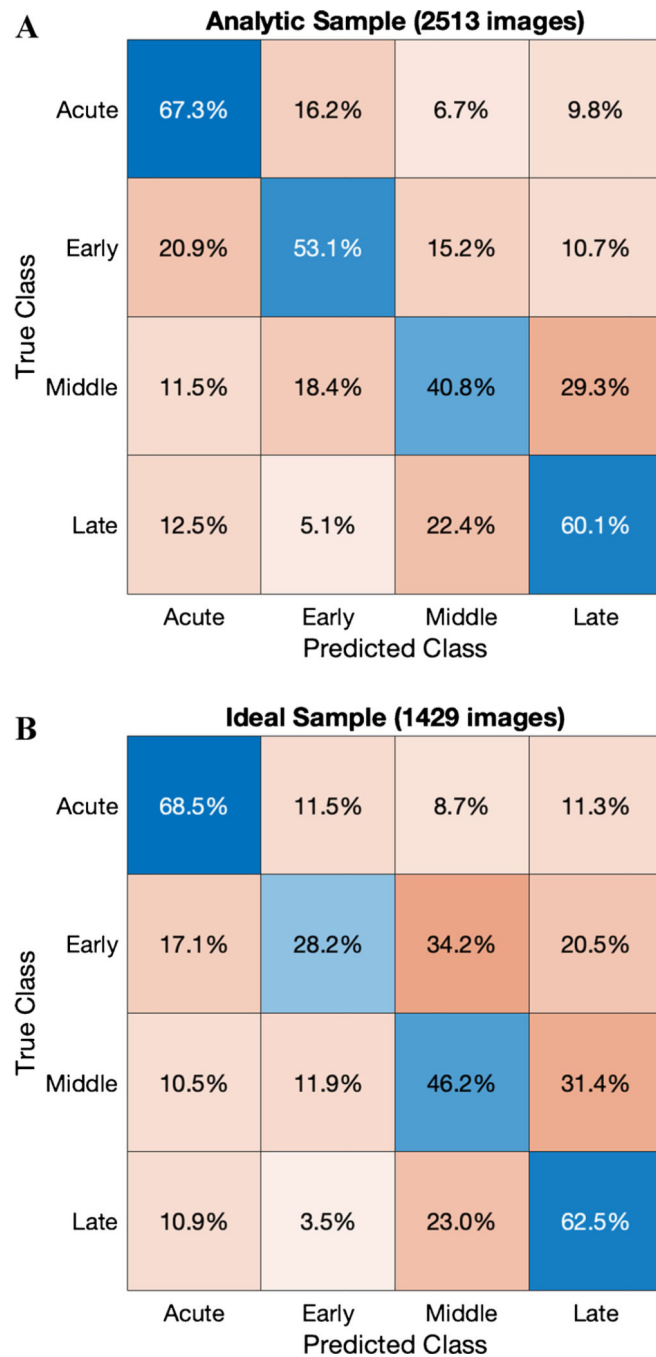


Fig. 4. Confusion matrices when classifying images into one of four fracture age classes for the analytic sample (a) and the ideal sample (b). Numbers are row-normalized and show the mean classification results across 5 repetitions of 5-fold cross-validation. Colors indicate correct (blue) and incorrect (red) classifications, with color-shades proportional to each number

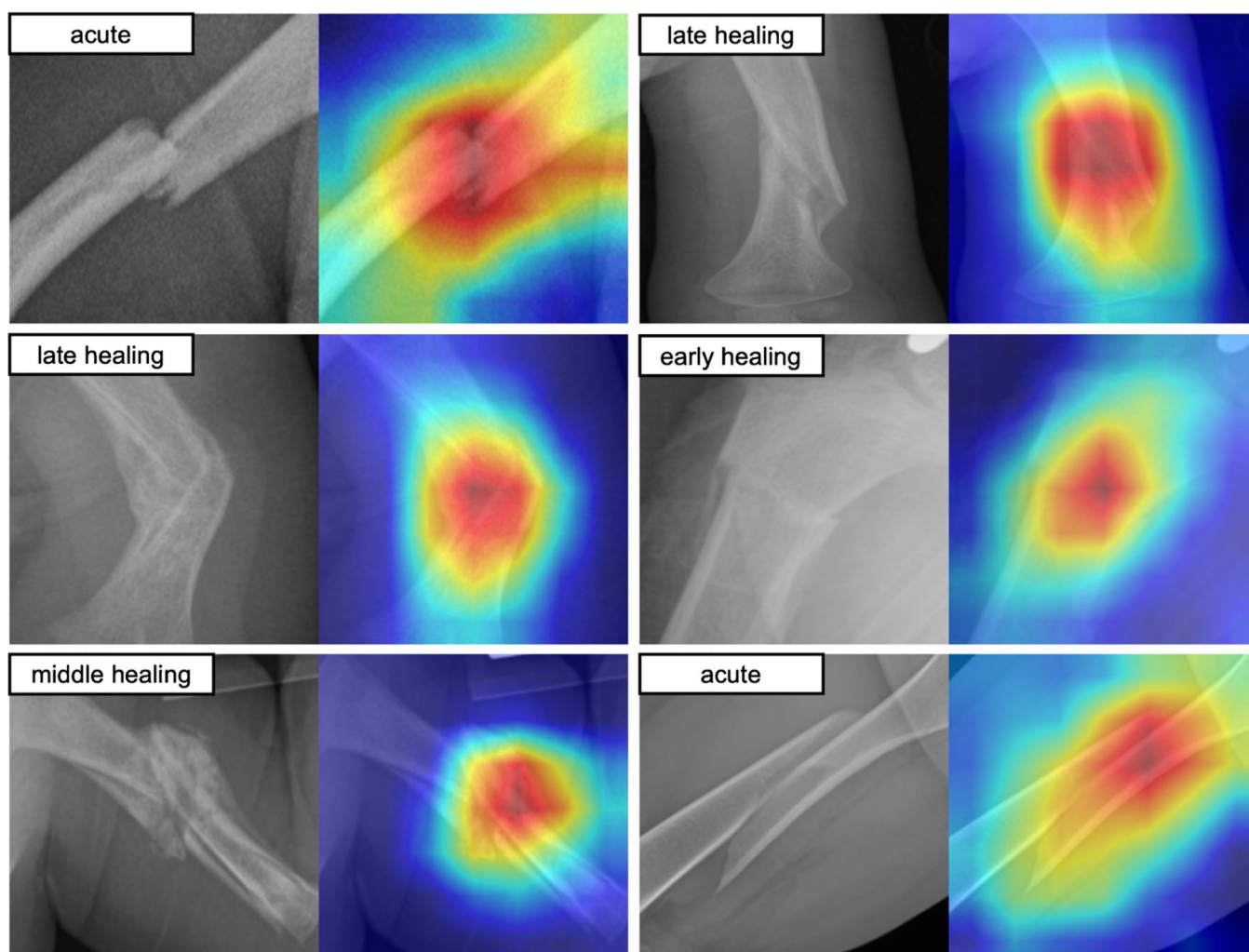


Fig. 5.
Example fractures and corresponding grad-CAM maps of a model trained on the ideal sample (all bones). *Red areas* suggest high impact on the model's classification decision while *blue areas* suggest low impact

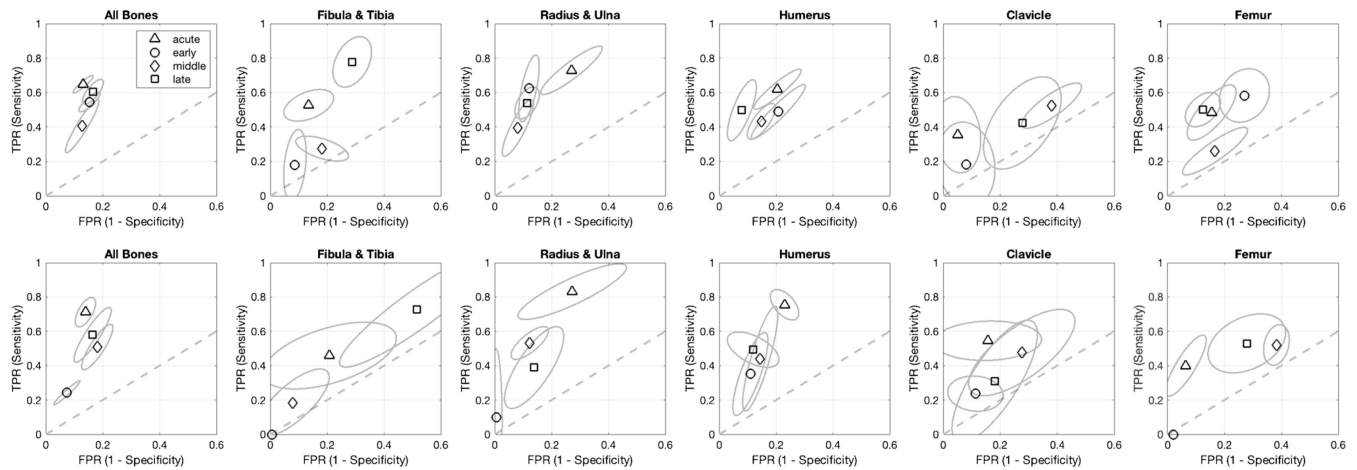


Fig. 6. Receiver operating characteristic (ROC) plots, showing sensitivity and specificity for each predicted fracture age class, with the full analytic sample on the top row and the ideal sample on the bottom row. *Ellipsoids* indicate 95% confidence areas based on the covariance matrix of different sensitivities and specificities across 5 repetitions of model training

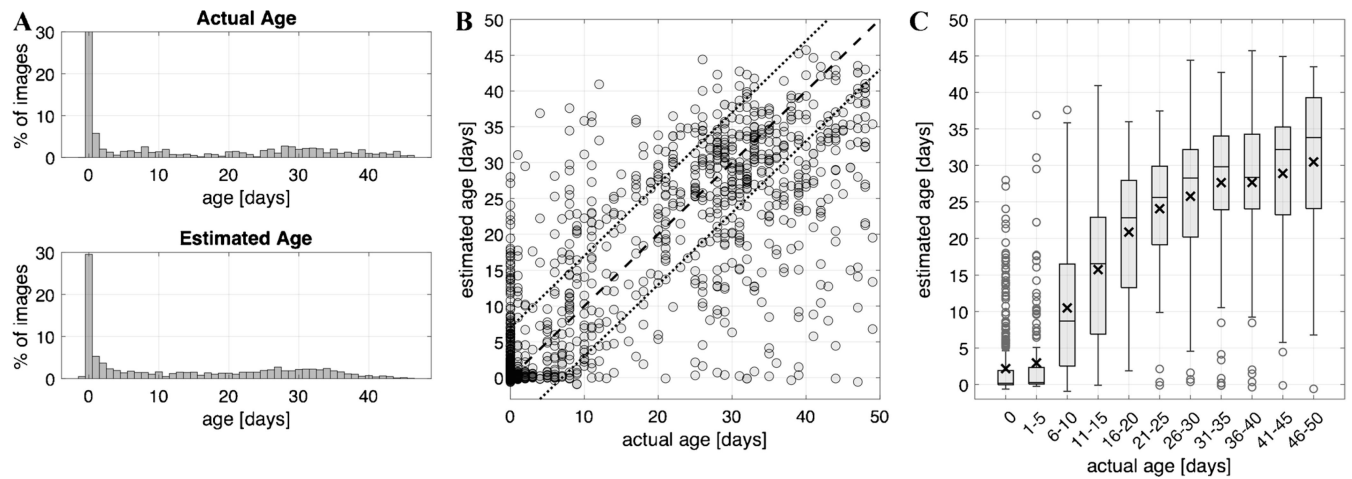


Fig. 7.

a Fracture age distribution of the ideal sample, limited to 0–49 days (1,246 images), with the actual distribution at the bottom and the estimated distribution at the top. **b** Scatterplot of actual age versus predicted age, with *dotted lines* demarcating absolute estimation errors within 7 days. *Circles* are opaque to better illustrate clusters of overlaying data points. The Pearson correlation coefficient is 0.80. **c** Box plots illustrating the distributions of predicted fracture age for different fracture age groups. The mean prediction for each age group is marked with an “ χ ”. Results represent an example of one cross-validated model trained with MAE loss

Table 1

Summary statistics of our bone fracture data, both for the analytics sample and the ideal sample. Patient age is based on the date of the first imaging. *std dev*, standard deviation. Note that some patients contributed fractures to multiple bones (radius and ulna)

Bone	Analytics sample					Ideal sample				
	No. of patients	No. of ROI images	No. of ROI images with casting	No. of ROI images with overlap	Mean no. of images per patient	Mean age of patient (years)	std dev age of patient (years)	No. of patients	No. of ROI images	Mean no. of images per patient
All bones	399	2,513	544	540	6.2	2.6	1.6	396	1,429	3.6
Clavicle	69	320	0	120	4.6	2.8	1.7	65	200	3.1
Femur	56	367	161	1	6.6	2.2	1.5	55	205	3.7
Fibula	31	159	20	61	5.1	3.3	1.8	27	78	3.7
Humerus	90	568	59	45	6.3	2.5	1.8	90	464	5.2
Radius	77	440	125	114	5.7	2.6	1.5	77	201	2.6
Tibia	60	311	78	108	5.2	2.9	1.7	53	125	2.4
Ulna	66	348	101	91	5.3	2.5	1.4	63	156	2.5

Table 2

Per class sensitivity and specificity when training and validating models on various data subsets. Parentheses indicate 95% CIs across 5 repetitions of 5-fold cross-validation

	Acute			Early healing			Middle healing			Late healing		
	No. of images	Sensitivity	Specificity	No. of images	Sensitivity	Specificity	No. of images	Sensitivity	Specificity	No. of images	Sensitivity	Specificity
Analytic sample												
All bones	1,079	0.65 (0.62, 0.67)	0.87 (0.85, 0.89)	335	0.55 (0.52, 0.57)	0.85 (0.83, 0.87)	706	0.41 (0.33, 0.48)	0.87 (0.84, 0.90)	393	0.60 (0.57, 0.64)	0.83 (0.82, 0.85)
Fibula and tibia	192	0.53 (0.48, 0.57)	0.86 (0.82, 0.91)	37	0.18 (0.07, 0.28)	0.91 (0.89, 0.93)	131	0.27 (0.24, 0.31)	0.82 (0.77, 0.86)	110	0.78 (0.70, 0.85)	0.71 (0.68, 0.75)
Radius and ulna	356	0.73 (0.66, 0.80)	0.73 (0.68, 0.78)	97	0.62 (0.53, 0.72)	0.88 (0.86, 0.90)	255	0.39 (0.31, 0.48)	0.92 (0.89, 0.95)	80	0.54 (0.49, 0.59)	0.89 (0.86, 0.91)
Humerus	242	0.62 (0.56, 0.68)	0.80 (0.75, 0.84)	94	0.49 (0.41, 0.57)	0.79 (0.74, 0.84)	150	0.43 (0.38, 0.49)	0.85 (0.83, 0.88)	82	0.50 (0.41, 0.59)	0.92 (0.89, 0.95)
Clavicle	144	0.36 (0.24, 0.47)	0.95 (0.91, 0.99)	47	0.18 (0.06, 0.31)	0.92 (0.87, 0.97)	87	0.52 (0.46, 0.59)	0.62 (0.57, 0.67)	42	0.42 (0.29, 0.56)	0.72 (0.65, 0.79)
Femur	145	0.48 (0.40, 0.57)	0.85 (0.80, 0.89)	60	0.58 (0.50, 0.66)	0.73 (0.69, 0.77)	83	0.26 (0.19, 0.33)	0.83 (0.77, 0.89)	79	0.50 (0.45, 0.55)	0.88 (0.84, 0.91)
Ideal sample												
All bones	645	0.71 (0.67, 0.76)	0.86 (0.84, 0.88)	117	0.24 (0.21, 0.28)	0.93 (0.90, 0.95)	411	0.51 (0.44, 0.58)	0.82 (0.79, 0.85)	256	0.58 (0.50, 0.66)	0.84 (0.80, 0.87)
Fibula and tibia	85	0.46 (0.36, 0.56)	0.79 (0.67, 0.91)	10	0.00 (0.00, 0.00)	0.99 (0.99, 1.00)	51	0.18 (0.08, 0.28)	0.92 (0.86, 0.99)	57	0.73 (0.56, 0.90)	0.48 (0.35, 0.62)
Radius and ulna	178	0.83 (0.75, 0.91)	0.73 (0.63, 0.82)	8	0.10 (-0.10, 0.30)	0.99 (0.98, 1.00)	133	0.53 (0.48, 0.58)	0.88 (0.84, 0.91)	38	0.39 (0.27, 0.51)	0.86 (0.81, 0.91)
Humerus	190	0.75 (0.71, 0.80)	0.77 (0.74, 0.80)	58	0.77 (0.74, 0.80)	0.89 (0.86, 0.92)	137	0.44 (0.29, 0.60)	0.86 (0.82, 0.89)	79	0.49 (0.44, 0.55)	0.88 (0.83, 0.93)
Clavicle	99	0.55 (0.49, 0.60)	0.84 (0.75, 0.94)	32	0.24 (0.19, 0.29)	0.89 (0.84, 0.94)	47	0.48 (0.35, 0.60)	0.72 (0.63, 0.82)	22	0.31 (0.13, 0.49)	0.82 (0.74, 0.90)
Femur	93	0.40 (0.31, 0.49)	0.94 (0.90, 0.97)	9	0.00 (0.00, 0.00)	0.98 (0.97, 0.99)	43	0.52 (0.46, 0.58)	0.62 (0.59, 0.64)	60	0.53 (0.44, 0.62)	0.72 (0.65, 0.79)

Table 3

MAE and RMSE (in days) when directly estimating fracture age and training/validating models on various data subsets. Parentheses indicate 95% CIs across 5 repetitions of 5-fold cross-validation. Note that the baseline error for models trained/evaluated on MAE is always guessing the median age of the training dataset while the baseline error for models trained/evaluated on MSE is guessing the mean age

Trained on MAE loss									
	Analytic sample (0–49 days)				Ideal sample (0–49 days)				
	No. of images	MAE (days)	Baseline error (median guess)	% within 7 days	No. of images	MAE (days)	Baseline error (median guess)	% within 7 days	
All Bones	2,258	6.81 (6.63, 7.00)	13.20	64.11 (62.79, 65.43)	1,246	6.17 (5.95, 6.40)	13.93	67.95 (66.66, 69.23)	
Fibula and tibia	407	9.07 (8.53, 9.61)	15.22	56.41 (53.77, 59.06)	174	11.73 (11.19, 12.27)	16.68	50.11 (45.82, 54.41)	
Radius and ulna	751	7.67 (7.05, 8.29)	13.50	62.90 (59.75, 66.06)	337	6.36 (5.94, 6.78)	15.22	67.72 (66.08, 69.35)	
Humerus	499	7.00 (6.57, 7.43)	11.14	64.13 (61.76, 66.50)	395	6.80 (6.47, 7.13)	12.20	65.42 (64.41, 66.43)	
Clavicle	291	7.44 (7.16, 7.71)	12.62	63.85 (61.60, 66.09)	186	7.85 (7.52, 8.17)	11.82	67.42 (66.12, 68.72)	
Femur	310	9.85 (9.55, 10.16)	13.62	56.58 (53.80, 59.36)	154	10.42 (10.18, 10.67)	13.47	57.14 (54.59, 59.69)	
Trained on MSE loss									
	Analytic sample (0–49 days)				Ideal sample (0–49 days)				
	No. of images	RMSE (days)	Baseline error (mean guess)	% within 7 days	No. of images	RMSE (days)	Baseline error (mean guess)	% within 7 days	
All Bones	2,258	9.79 (9.48, 10.09)	15.34	63.60 (62.79, 64.42)	1,246	9.42 (9.31, 9.53)	15.95	68.30 (67.14, 69.45)	
Fibula and tibia	407	13.17 (12.86, 13.46)	17.00	52.09 (50.35, 53.83)	174	16.31 (15.64, 16.96)	18.16	42.76 (38.46, 47.06)	
Radius and ulna	751	11.73 (11.04, 12.38)	15.32	53.13 (47.66, 58.60)	337	10.96 (10.64, 11.27)	16.66	61.13 (58.98, 63.28)	
Humerus	499	9.67 (9.54, 9.80)	13.32	62.48 (59.69, 65.28)	395	10.59 (10.14, 11.03)	14.10	60.86 (59.23, 62.49)	
Clavicle	291	11.16 (10.94, 11.37)	14.93	58.56 (56.37, 60.74)	186	12.51 (12.01, 12.99)	14.63	54.30 (52.60, 56.00)	
Femur	310	14.79 (14.28, 15.29)	16.14	48.58 (43.76, 53.40)	154	16.85 (16.49, 17.20)	17.14	50.78 (47.02, 54.54)	