Published in final edited form as:

*Pediatr Radiol*. 2023 May ; 53(6): 1117–1124. doi:10.1007/s00247-023-05590-0.

# Dating birth-related clavicular fractures: pediatric radiologists versus artificial intelligence

**Andy Tsai**[1], **Jeannette M. Pérez-Rosselló**[1], **Kirsten Ecklund**[1], **Michele M. Walters**[1], **Paul K. Kleinman**[1]

[1]Department of Radiology, Boston Children's Hospital, Harvard Medical School, 300 Longwood Ave., Boston, MA 02115, USA

## Abstract

**Background—**Fracture dating from skeletal surveys is crucial in the diagnosis and investigation of infant abuse. However, this task is challenging because of the subjective nature of the radiologic interpretation and the lack of ground truth. Researchers have used birth-related clavicle fractures as a surrogate to study the radiographic pattern of healing; however, they did not elucidate the accuracy performance of the radiologists in dating fractures.

**Objective—**To determine the accuracy of radiologists in dating birth-related clavicle fractures and compare their performance to that achieved by computer algorithm.

**Materials and methods—**We used a previously assembled birth-related clavicle fracture database consisting of 416 anteroposterior clavicle radiographs as the study cohort. The average and standard deviation of the fracture age within this database were 24 days and 18 days, respectively. Three blinded radiologists independently estimated the ages of the clavicle fractures depicted in the radiographs within the database. We compared these estimation results to those made by a recently published deep-learning (DL) model conducted with the identical infant cohort. We calculated standard error metrics to compare the accuracy performances of the radiologists and the computer model.

**Results—**The intra- and inter-reader agreements of the fracture age estimates by the radiologists were moderate to good. The radiologists estimated the fracture ages with a mean absolute error (MAE) of 6.1–7.1 days, and standard deviation of the absolute error of 6.3–8.3 days. The accuracy performances of the three radiologists were not significantly different from one another. In comparison, the DL model estimated the age of clavicle fractures with an MAE of 4.2 days, significantly lower than all of the radiologists ($P < 0.001$).

**Conclusion—**Three experienced pediatric radiologists dated clavicular fractures with moderate–good intra- and inter-reader agreements. The correlations between the radiologists' estimates and the ground truth were moderate to good. The fracture ages assigned by the DL model showed superior correlation with the ground truth compared to radiologists' dating estimates.

✉Andy Tsai, andy.tsai@childrens.harvard.edu.

**Conflicts of interest** None

**Keywords**

Artificial intelligence; Child abuse; Clavicle; Deep learning; Fracture; Fracture dating; Healing; Infants; Radiography

## Introduction

Temporal patterns of fracture healing in childhood are typically described in the literature in general terms, and these rough timelines are usually sufficient to guide clinical management. However, in the context of suspected infant abuse, more precise fracture dating is desirable. Dating can establish a timeline that confirms the presence of fractures of different ages — a hallmark of infant abuse. Also, child protection agencies and law enforcement often rely upon radiologists' opinion of fracture age to support or refute a purported timeline of injury. This can enable the authorities to address any potential risk to children in the environment and prosecute alleged perpetrators of the abuse.

A number of studies have developed schemes to date fractures based on their radiographic features [1–10]. Most have analyzed radiographs in cases of accidental fractures when the date of injury was known [1–5, 7, 9, 10]. The use of cohorts of abused infants where the ground truth is determined by witness descriptions or perpetrator admissions is potentially problematic [6, 8]. A recent study explored the use of machine learning models to date childhood/adolescent fractures [11]. In that study, the fractures were from a wide variety of anatomical regions in individuals 0–19 years old. Only 10 infant fractures were included in the cohort. To estimate the fracture age, the study proposed two machine learning models and a fracture healing scale. All three of these methodologies generated similar estimates with substantial imprecision. Specifically, the standard deviations of the mean differences for all tested models ranged from a low of 52 days for the displaced fractures to a high of 82 days for buckle fractures [11].

Birth-related clavicle fractures have been used by several authors as a surrogate to establish the evolution of radiographic patterns of fracture healing in cases of infant abuse [12, 13]. These investigators provided the following rationale for using the clavicle as a surrogate: (1) the fracture age is known precisely (i.e. occurring at birth), (2) clavicular morphology is similar to that of other tubular bones and thus its healing pattern is likely to be generalizable to those fractures involving the long bones and ribs — other common sites of inflicted injury, and (3) infant clavicular fractures are not immobilized and thus the healing pattern should be similar to that of abusive fractures in which treatment is often delayed. These clinical studies yielded reasonable guidance for dating fractures based on the radiographic features, but they indicated little about how accurately radiologists might perform in dating fractures in the clinical setting. To our knowledge, no published studies have evaluated the performance of radiologists in the dating of clavicular fractures in infants.

In the current study, we aimed to determine the accuracy of the radiologists' age estimates of birth-related clavicular fractures. Furthermore, we compared the radiologists' performance to the published data that employed an artificial intelligence algorithm utilizing a deep learning (DL) model derived from the same cohort of infant clavicular fractures [14].

## Materials and methods

Our institutional review board approved this single-institution retrospective study and waived the patient informed consent. This study complied with the Health Insurance Portability and Accountability Act guidelines.

### Study sample

For this investigation, we used the birth-related clavicle fracture database assembled by Tsai et al. [14]. The radiographs in this database were obtained at a large tertiary children's hospital (July 2003–March 2022). Straight anteroposterior (AP) and 15° cephalad angulated AP views of the clavicle were included in the database, and we considered each view as a distinct image sample and standalone image of the clavicle. The inclusion criteria were: (1) infants ≤ 3 months old; (2) confirmed birth-related clavicle fracture; (3) infants without significant health concerns (e.g., congenital heart disease or sepsis that could alter the pattern of fracture healing), skeletal dysplasia or metabolic bone disease; and (4) adequate image quality and field of view. Given that the clavicle fractures occurred at birth, each patient's chronological age at the time of the radiograph was assigned as the ground truth fracture age.

The curated database consisted of 416 dedicated clavicle radiographs (right = 218, left = 198) from 213 infants (male = 116, female = 97). The average and standard deviation of the fracture age were 24 days and 18 days, respectively; and the ages ranged 2–88 days. Radiographically, clavicle fractures ≥ 60 days old are difficult to distinguish from one another, and thus all clavicle fractures ≥ 60 days old were assigned to the same age rank, i.e. 60 + days. With this re-categorization, the fracture age within the database ranged 2–60 + days (total of 59 age ranks).

### Fracture age estimation by radiologists

Three board-certified radiologists (K.E. with 26 years, J.M.P.-R. with 20 years and M.W. with 13 years of post-fellowship experience in pediatric musculoskeletal imaging) were blinded to study design, patient demographics and radiograph results. They independently estimated the ages of the clavicle fractures shown on the radiographs using the picture archiving and communication system. The radiologists were told a priori that there were 59 fracture age ranks within the database, ranging from 2 days to 60 + days. They were asked to provide the age estimates in number of days. To assess intra-reader variability, we randomly selected 25% of the clavicle fractures within the database and the three radiologists re-estimated their fracture ages > 1 month later.

### Statistical analyses

Reader agreement was assessed via intra-class correlation (ICC) [15]. An accepted guideline for the interpretation of ICC is to consider values < 0.50 as poor, 0.50–0.75 as moderate, 0.75–0.90 as good and 0.90–1.00 as excellent correlations [16]. To assess the accuracy performance of the three readers and the DL model, we employed standard error metrics including mean absolute error (MAE), root mean square error (RMSE), standard deviation of absolute error (SDAE) and cumulative score (CS). Both MAE and RMSE range from 0 to infinity, with values closer to 0 indicating more accurate estimates. SDAE also ranges from 0 to infinity, with small values indicating tighter clustering of the estimates. The cumulative score $CS_j$ was defined as $(N_j/N) \times 100\%$, where N was the total number of clavicle radiographs in the database and $N_j$ was the number of clavicle radiographs with estimated fracture ages within j days from the ground truth. For example, $CS_7$ is the percentage of age estimates within 7 days of the ground truth. We also generated Bland–Altman plots and scatterplots of the estimated versus the ground truth fracture ages to visually compare the performances of the different approaches to fracture dating.

### Comparison between radiologists and deep learning model

We compared the performance data of the radiologists to the previously published data of fracture age estimation employing a DL model [14]. Our case selection, methodology, final case cohort and error metric analyses were identical to those of this prior publication. We performed the pairwise statistical comparisons of the absolute estimation errors between the three readers and the DL model using Wilcoxon signed rank test. To determine whether the accuracy performance of the three radiologists and the DL model were a function of fracture age, we partitioned the database into three consecutive non-overlapping fracture age groups and then calculated the estimation error metrics (MAE, RMSE and SDAE) of each group for the three radiologists and the DL model. To assess statistical significance of the estimation errors within each of the three non-overlapping fracture age groups, we conducted pairwise comparisons of the absolute errors of the three radiologists and the DL model via the Wilcoxon signed rank test. All analyses were conducted with a statistical significance of $P < 0.001$ using MATLAB 8.3 (MathWorks, Natick, MA).

## Results

Figures 1, 2, 3, 4 and 5 show example radiographs of birth-related clavicle fractures and the corresponding age estimates by the three radiologists and the DL model [14]. These five examples depict the wide radiographic spectrum of the clavicle fractures as a function of fracture age and illustrate the difficulties in providing accurate fracture age estimates by the radiologists and the DL model.

### Fracture age estimation by radiologists

For the fracture age estimates made by the three radiologists, the intra-reader agreements were moderate to good (ICC range 0.62 to 0.82), and the inter-reader agreements were moderate to good (ICC range 0.66 to 0.81). The correlation between the radiologists' estimates and the ground truth were moderate to good (ICC range 0.69 to 0.84). The mean of the absolute errors by the three radiologists was 6.1–7.1 days (with standard deviation of 6.3–8.3 days). Pairwise statistical comparisons of the radiologists' absolute errors showed that they were not significantly different from one another ($P$-values ranged from 0.20 to 0.78). The RMSE of the age estimates by the radiologists ranged from 8.8 to 10.9 days. The percentage of estimates within 0 days of the ground truth ($CS_0$; in other words, the accuracy percentage of the estimates) ranged between 8.7% and 9.6%. The percentage of estimates within 7 days of the ground truth ($CS_7$) ranged between 69.2% and 71.4%. Additional details regarding the accuracy performances of the three radiologists are listed in Table 1.

The scatterplots and the Bland–Altman plots visually compare the age estimates by the radiologists to the ground truth and are shown in Figs. 6 and 7. Based on the Bland–Altman plots, the absolute biases of the estimates by the three radiologists were 1.8, 5.6 and 3.7 days; and the standard deviations were 8.6, 9.4 and 9.7 days, respectively.

### Comparison between radiologist and deep learning model data

The age estimates made by the DL model [14] showed superior correlation with the ground truth as compared to the age estimates made by the radiologists (0.92 vs. 0.69–0.84). Pairwise statistical comparisons showed that the absolute errors made by the DL model were significantly less than those made by any of the three radiologists ($P < 0.001$). Specifically, the MAE of the DL model was 4.2 days, which was 31–41% less than that of the radiologists. Cumulative score curves show that the DL model performed better than all of the radiologists at every increment of the absolute error, from 0 to 14 days (Fig. 8).

The scatterplot of the estimated versus ground truth fracture ages demonstrates a general tendency of increasing estimation error with increasing fracture age (Fig. 6). Table 2 shows the findings when the true fracture ages were empirically partitioned into non-overlapping age groups (< 15 days old [$n = 15$], ≥ 15 to < 30 days old [$n = 13$], and ≥ 30 days old [$n = 129$]). For the youngest and the middle age groups, reader 1 had significantly larger absolute errors than the other radiologists and the DL model ($P < 0.001$). For the oldest age group, readers 2 and 3 had significantly larger absolute errors than reader 1 and the DL model ($P < 0.001$). These two readers tended to underestimate the fracture ages in the oldest age group. Overall, based on MAE, RMSE and SDAE, the DL model was the most accurate in each of the three age groups.

## Discussion

In this study, we compared the accuracy performances of pediatric radiologists with a DL model in estimating the ages of birth-related clavicle fractures visible on radiographs. The radiologists estimated the fracture ages with moderate-to-good intra- and inter-reader agreements — achieving an MAE of 6.1–7.1 days and $CS_7$ of 69.2–71.4%. In comparison, the DL model estimated the ages of the clavicle fractures with an MAE of 4.2 days, with significantly smaller error and higher cumulative score than all three radiologists. These differences in accuracy would have likely been greater if general radiologists were tested because all of our readers were quite experienced pediatric radiologists — each with subspecialty expertise in musculoskeletal imaging. Radiologists were most accurate in fracture dating during the early phases of healing, when subperiosteal new bone formation and bony callus become evident. The accuracy performances of the radiologists deteriorated with increasing fracture age, when these features of fracture healing gradually evolve and mature.

Past studies that used birth-related clavicle fractures were performed with the reasonable expectation that dating based on the appearance and evolution of subperiosteal new bone formation and callus would be generalizable to other tubular bones [12, 13]. In contrast, the black-box nature of DL means that we are unable to discern why the algorithm chooses to assign a particular age to a given fracture, so our site-specific DL algorithm would not be generalizable to fractures at other anatomical locations.

How would implementation of this DL approach to dating clavicle fractures impact a radiologist's interpretation of infant skeletal surveys for suspected abuse? In a study by Barber et al. [17] of 567 infants undergoing skeletal surveys for suspected abuse, clavicular fractures were encountered in 4% of cases — a modest but not inconsequential number given that the skeletal surveys were positive for at least one unsuspected fracture in only 119/567 (21%) infants [17]. Additionally, given that clavicle fractures are often incidentally noted on a chest radiograph, or are imaged because a suspected clavicular fracture is palpated in a young infant, accurate dating often helps answer the question of whether an infant's clavicular fracture is consistent with a birth injury. The value of radiologic dating with DL is likely to grow substantially if this approach is successfully applied to injuries at other anatomical sites — especially the ribs and long bones.

To our knowledge, this is the first published study that systematically examined the performance of pediatric radiologists in dating infant fractures. It confirms the generally held belief that, based on their experience and knowledge, pediatric radiologists can reliably date certain fractures, and those estimates have a firm place in clinical and medicolegal environments. Our study also indicates that a DL algorithm can inform a radiologist's assessment, leading to more accurate fracture age estimates. Investigators should be encouraged to confirm our results and to leverage our methodology in the study of fractures at other anatomical locations where a ground truth can be firmly established. Toward this end, we plan to study an infant cohort consisting of radiographs with iatrogenically induced rib fractures occurring at the time of thoracotomy for short-gap esophageal atresia, congenital heart disease and various pulmonary/mediastinal lesions. Furthermore, we are conducting a review of radiographs of humeral and femoral fractures occurring in the obstetrical setting. The numbers might be insufficient to replicate the cohort size of the current study and might require multicenter collaboration. These investigations should broaden the evidence base that confirms radiologists' ability to reliably date fractures in young infants and are likely to expand and solidify the role of DL as a virtual consultant to the interpreting radiologist.

This study has few limitations. First, the clavicle fracture database was curated from a single institution by a sole radiologist, which might have resulted in selection bias. A large and diverse database drawing from multiple institutions, vetted by multiple radiologists, would minimize this type of bias. To further generalize our study findings, we could employ an external dataset for validation. Last, our retrospective study might not represent the true performance levels of the radiologists or the DL model. Specifically, the manual cropping and centering of the clavicle fractures for radiologists' interpretation and DL model estimation do not reflect the actual clinical environment. Future investigations employing prospective methodology would be of value to mitigate some of these limitations and to validate the findings of this encouraging preliminary study.

## Conclusion

Three experienced pediatric radiologists dated birth-related clavicular fractures with moderate–good intra- and inter-reader agreements. The correlations between the radiologists' estimates and the ground truth were moderate to good. The DL model estimates showed even better correlation with the ground truth. This study demonstrates that pediatric radiologists perform quite well in dating infant clavicular fractures with implications for dating abusive injuries. Furthermore, it supports a promising role for DL as a virtual consultant to the interpreting radiologist in cases of suspected infant abuse.

## Acknowledgements

Research reported in this publication was supported in part by the Eunice Kennedy Shriver National Institute of Child Health and Human Development of the National Institutes of Health under award number R21HD108634 to Andy Tsai. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health. Research reported in this publication was also supported in part by the Radiological Society of North America (RSNA) Research & Education Foundation, through Research Seed Grant RSD2133 to Andy Tsai. The content of this publication is solely the responsibility of the authors and does not necessarily represent the official views of the RSNA.

## References

1. Cumming WA (1979) Neonatal skeletal fractures. Birth trauma or child abuse? J Can Assoc Radiol 30:30–33 [PubMed: 429433]
2. O'Connor JF, Cohen J (1987) Dating fractures. In: Kleinman PK (ed) Diagnostic imaging of child abuse. Williams & Wilkins, Baltimore, pp 103–113
3. Yeo LI, Reed MH (1994) Staging of healing of femoral fractures in children. Can Assoc Radiol J 45:16–19 [PubMed: 8118709]
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 175:75–78 [PubMed: 10882250]
5. Malone CA, Sauer NJ, Fenton TW (2011) A radiographic assessment of pediatric fracture healing and time since injury. J Forensic Sci 56:1123–1130 [PubMed: 21644992]
6. Halliday K, Broderick N, Somers J, Hawkes R (2011) Dating fractures in infants. Clin Radiol 66:1049–1054 [PubMed: 21763645]
7. Prosser I, Lawon Z, Evans A et al. (2012) A timetable for the radiologic features of fracture healing in young children. AJR Am J Roentgenol 198:1014–1120 [PubMed: 22528890]
8. Sanchez TR, Nguyen H, Palacios W et al. (2013) Retrospective evaluation and dating of non-accidental rib fractures in infants. Clin Radiol 68:e467–e471 [PubMed: 23622800]
9. 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 46:333–341 [PubMed: 28070625]
10. 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]
11. Kyllonen KM, Monson KL, Smith MA (2022) Postmortem and antemortem forensic assessment of pediatric fracture healing from radiographs and machine learning classification. Biology 11:749 [PubMed: 35625477]
12. Walters MM, Forbes PW, Buonomo C, Kleinman PK (2014) Healing patterns of clavicular birth injuries as a guide to fracture dating in cases of possible infant abuse. Pediatr Radiol 44:1224–1229 [PubMed: 24777389]
13. Fadell M, Miller A, Trefan L et al. (2017) Radiological features of healing in newborn clavicular fractures. Eur Radiol 27:2180–2187 [PubMed: 27629420]
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 52:2206–2214 [PubMed: 35578043]

15. Koch GG (1982) Intraclass correlation coefficient. In: Kotz S, Johnson NL (eds) Encyclopedia of statistical sciences, vol 4. Wiley, New York, pp 213–217
16. Koo TK, Li MY (2016) A guideline of selecting and reporting intraclass correlation coefficients for reliability research. J Chiropr Med 15:155–163 [PubMed: 27330520]
17. Barber I, Perez-Rossello JM, Wilson CR, Kleinman PK (2015) The yield of high-detail radiographic skeletal surveys in suspected infant abuse. Pediatr Radiol 45:69–80 [PubMed: 24997790]

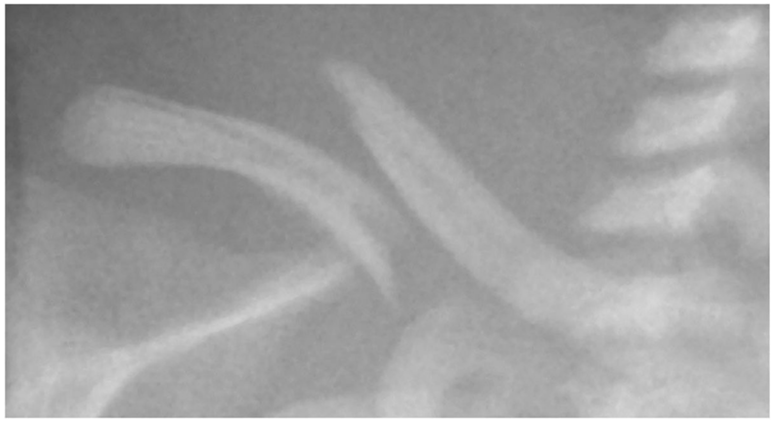

**Fig. 1.**

Anteroposterior clavicle radiograph of a 4-day-old boy with birth-related clavicle fracture. Readers 1, 2 and 3 estimated the age of the fracture as 3, 6 and 5 days, respectively. In comparison, the deep-learning (DL) model estimated the fracture age as 2 days [14]

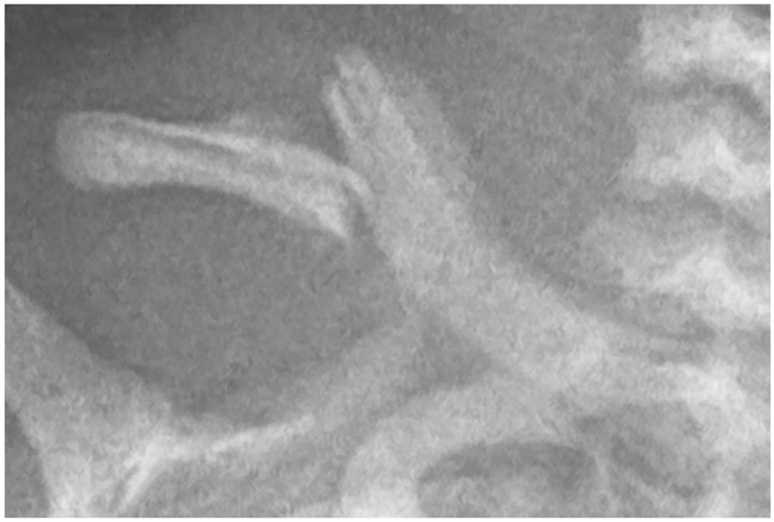

**Fig. 2.**

Anteroposterior clavicle radiograph of a 12-day-old girl with birth-related clavicle fracture. Readers 1, 2 and 3 estimated the age of the fracture as 14, 10 and 16 days, respectively. In comparison, the DL model estimated the fracture age as 13 days [14]

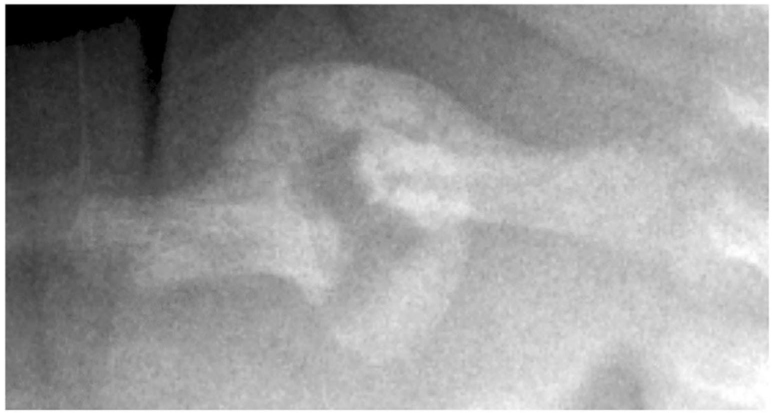

**Fig. 3.**
Anteroposterior clavicle radiograph of an 18-day-old girl with birth-related clavicle fracture. Readers 1, 2 and 3 estimated the age of the fracture as 28, 23 and 14 days, respectively. In comparison, the DL model estimated the fracture age as 19 days [14]

**Fig. 4.**

Anteroposterior clavicle radiograph of a 37-day-old boy with birth-related clavicle fracture. Readers 1, 2 and 3 estimated the age of the fracture as 35, 38 and 25 days, respectively. In comparison, the DL model estimated the fracture age as 28 days [14]

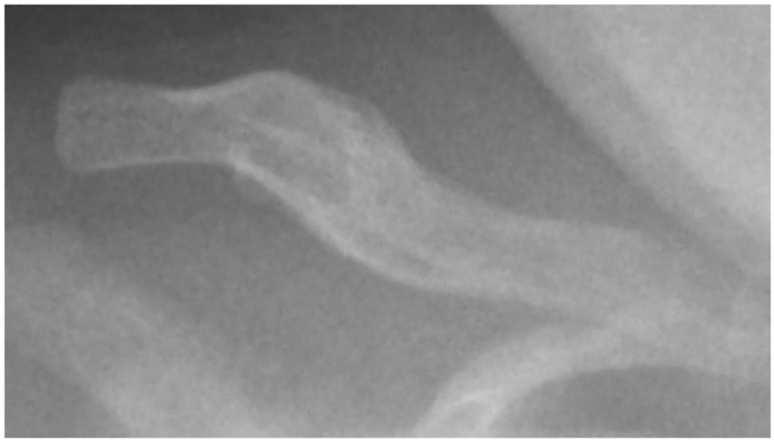

**Fig. 5.**
Anteroposterior clavicle radiograph of a 57-day-old boy with birth-related clavicle fracture. Readers 1, 2 and 3 estimated the age of the fracture as 60, 22 and 30 days, respectively. In comparison, the DL model estimated the fracture age as 60 days [14]

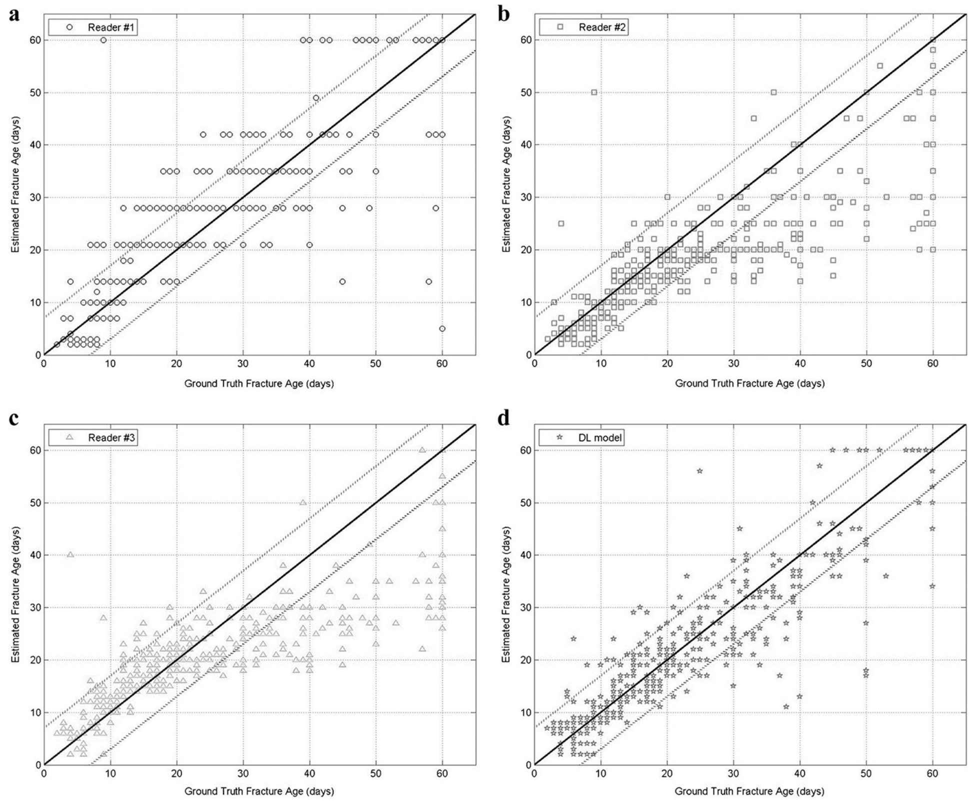


**Fig. 6.**
Scatterplots of estimated fracture age versus the ground truth to assess for estimation accuracy. The solid black line denotes the identity line (where the estimated age equals the ground truth). The black dashed lines demarcate where the absolute estimation error equals 7 days. **a** Reader 1. **b** Reader 2. **c** Reader 3. **d** Deep learning (DL) model [14]. In comparing these scatterplots, the estimates by the DL model have the least amount of bias and data dispersion. Each plot contains 416 data points

Author Manuscript

Author Manuscript

Author Manuscript



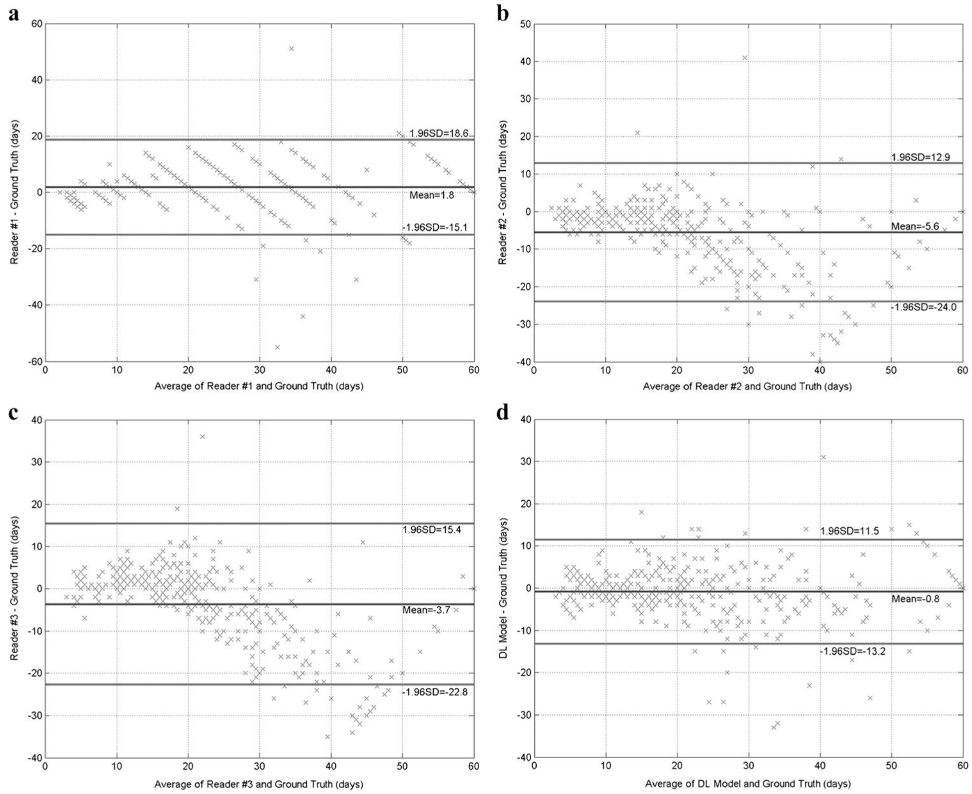


**Fig. 7.**

Bland–Altman plots of estimated fracture age and the ground truth to assess for estimation accuracy. **a** Plot of difference between reader 1 fracture age estimates and the ground truth shows bias of 1.8 days and standard deviation of 8.6 days. **b** Plot of difference between reader 2 fracture age estimates and the ground truth shows bias of −5.6 days and standard deviation of 9.4 days. **c** Plot of difference between reader 3 fracture age estimates and the ground truth shows bias of −3.7 days and standard deviation of 9.7 days. **d** Plot of difference between deep learning (DL) model fracture age estimates and the ground truth shows bias of −0.8 days and standard deviation of 6.3 days [14]. In comparing these Bland–Altman plots, the limits of agreement is the tightest for the DL model. Each plot contains 416 points

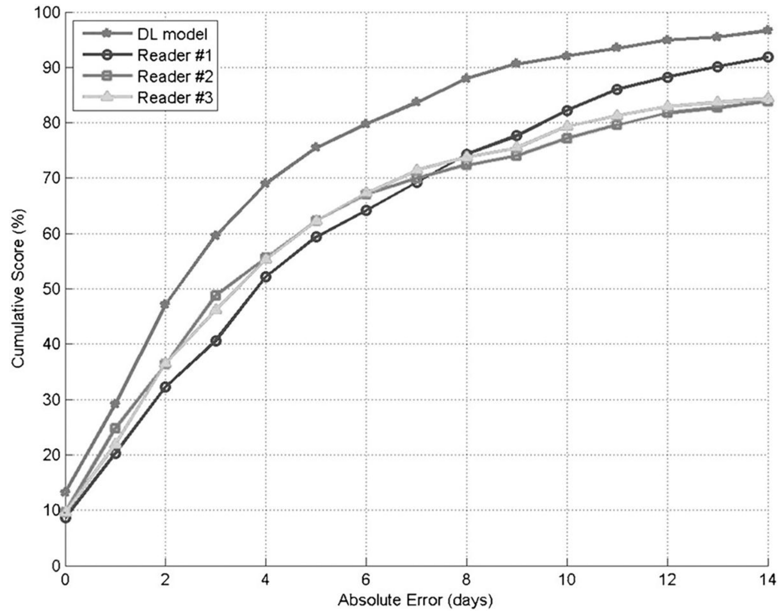


**Fig. 8.**

Cumulative score curves of the three radiologists and the deep learning (DL) model, at absolute error rates ranging from 0 to 14 days. The cumulative scores for the three radiologists are similar. In comparison, the cumulative score curve for the DL model [14] rises faster than the radiologists' curves and is more accurate at every increment of absolute error from 0 to 14 days

## Table 1

Accuracy performance of the three radiologists

| Error metrics | Reader 1 | Reader 2 | Reader 3 |
|---|---|---|---|
| MAE (days) | 6.1 | 7.1 | 6.9 |
| SDAE (days) | 6.3 | 8.3 | 7.8 |
| Maximum absolute error (days) | 55 | 41 | 36 |
| RMSE (days) | 8.8 | 10.9 | 10.4 |
| ICC | 0.84 | 0.70 | 0.69 |
| [95% CI] | [0.81, 0.87] | [0.64, 0.74] | [0.64, 0.74] |
| *P*-value | < 0.0001 | < 0.0001 | < 0.0001 |
| $CS_0$ (%) | 8.7 | 9.6 | 9.6 |
| $CS_1$ (%) | 20.2 | 24.8 | 21.9 |
| $CS_2$ (%) | 32.2 | 36.3 | 36.5 |
| $CS_3$ (%) | 40.6 | 48.8 | 46.2 |
| $CS_4$ (%) | 52.2 | 55.5 | 55.3 |
| $CS_5$ (%) | 59.4 | 62.3 | 62.3 |
| $CS_6$ (%) | 64.2 | 67.1 | 67.3 |
| $CS_7$ (%) | 69.2 | 70.0 | 71.4 |
| $CS_8$ (%) | 74.3 | 72.4 | 73.8 |
| $CS_9$ (%) | 77.6 | 74.0 | 75.5 |
| $CS_{10}$ (%) | 82.2 | 77.2 | 79.3 |
| $CS_{11}$ (%) | 86.1 | 79.6 | 81.3 |
| $CS_{12}$ (%) | 88.2 | 81.7 | 82.9 |
| $CS_{13}$ (%) | 90.1 | 82.7 | 83.7 |
| $CS_{14}$ (%) | 91.8 | 83.9 | 84.4 |

*CI* confidence interval, *CS* cumulative score, *ICC* intraclass correlation coefficient, *MAE* mean absolute error, *RMSE* root mean square error, *SDAE* standard deviation of absolute error

## Table 2

Accuracy performance of the radiologists and the deep-learning (DL) model as a function of fracture age

| | MAE (days) | | | SDAE (days) | | | RMSE (days) | | |
|---|---|---|---|---|---|---|---|---|---|
| **Fracture age range (days)** | **< 15** | **≥ 15 to < 30** | **≥ 30** | **< 15** | **≥ 15 to < 30** | **≥ 30** | **< 15** | **≥ 15 to < 30** | **≥ 30** |
| Reader 1 | 4.4 | 6.3 | 7.9 | 5.4 | 4.3 | 8.3 | 6.9 | 7.6 | 11.4 |
| Reader 2 | 2.7 | 4.1 | 15.4 | 3.9 | 3.0 | 9.8 | 4.7 | 5.1 | 18.2 |
| Reader 3 | 3.2 | 3.4 | 14.9 | 3.8 | 2.7 | 8.8 | 5.0 | 4.3 | 17.3 |
| DL model[a] | 2.7 | 3.7 | 6.3 | 2.6 | 3.8 | 6.6 | 3.7 | 5.3 | 9.1 |

*MAE* mean absolute error, *RMSE* root mean square error, *SDAE* standard deviation of absolute error

[a] Previously published data[14]