Deep learning for the radiographic diagnosis of proximal femur fractures: Limitations and programming issues.

Sylvain Guy, Christophe Jacquet, Damien Tsenkoff, Jean-Noël Argenson, Matthieu Ollivier

Original article

Deep learning for the radiographic diagnosis of proximal femur fractures: Limitations and programming issues

Sylvain Guy*, Christophe Jacquet, Damien Tsenkoff, Jean-Noël Argenson, Matthieu Ollivier

Institut du Mouvement et de l'appareil Locomoteur, 270, boulevard de Sainte Marguerite, 13009 Marseille, France

ARTICLE INFO

Keywords:
Deep learning
Artificial intelligence
Traumatology
Proximal femur fracture

ABSTRACT

Introduction: Radiology is one of the domains where artificial intelligence (AI) yields encouraging results, with diagnostic accuracy that approaches that of experienced radiologists and physicians. Diagnostic errors in traumatology are rare but can have serious functional consequences. Using AI as a radiological diagnostic aid may be beneficial in the emergency room. Thus, an effective, low-cost software that helps with making radiographic diagnoses would be a relevant tool for current clinical practice, although this concept has rarely been evaluated in orthopedics for proximal femur fractures (PFF). This led us to conduct a prospective study with the goals of: 1) programming deep learning software to help make the diagnosis of PFF on radiographs and 2) to evaluate its performance.
Hypothesis: It is possible to program an effective deep learning software to help make the diagnosis of PFF based on a limited number of radiographs.
Methods: Our database consisted of 1309 radiographs: 963 had a PFF, while 346 did not. The sample size was increased 8-fold (resulting in 10,472 radiographs) using a validated technique. Each radiograph was evaluated by an orthopedic surgeon using RectLabel™ software (https://rectlabel.com), by differentiating between healthy and fractured zones. Fractures were classified according to the AO system. The deep learning algorithm was programmed on Tensorflow™ software (Google Brain, Santa Clara, Ca, USA, tensorflow.org). In all, 9425 annotated radiographs (90%) were used for the training phase and 1074 (10%) for the test phase.
Results: The sensitivity of the algorithm was 61% for femoral neck fractures and 67% for trochanteric fractures. The specificity was 67% and 69%, the positive predictive value was 55% and 56%, while the negative predictive value was 74% and 78%, respectively.
Conclusion: Our results are not good enough for our algorithm to be used in current clinical practice. Programming of deep learning software with sufficient diagnostic accuracy can only be done with several tens of thousands of radiographs, or by using transfer learning.
Level of evidence: III; Diagnostic studies, Study of nonconsecutive patients, without consistently applied reference "gold" standard.

1. Introduction

Artificial intelligence (AI) plays an important role in our day-to-day lives [1,2]. When compared to human intelligence, it can hold its own against professional video gamers [3], poker players [4] or even Go players [5]. Such good performances can be explained by the development of "deep learning" [6], a learning method based on human convolutional neural networks.

The medical world has also been impacted by AI [7], especially molecular biology [8,9], ophthalmology [10], dermatology [11], and even anatomical pathology [12,13]. However, the progress has been the most striking in radiology [14–17]. Despite inconclusive results early on [18], the rapid development of deep learning has allowed AI to approach human performance. Thus it comes close to the performance of trained radiologists in reading mammograms [19], and can help to optimize the radiologist's work by doing triage before the analysis [20]. However, no program has successfully surpassed

Abbreviations: AI, Artificial Intelligence; PFF, proximal femur fracture; ED, emergency department.

* Corresponding author.
E-mail address: sylvain.guy.vidal@gmail.com (S. Guy).

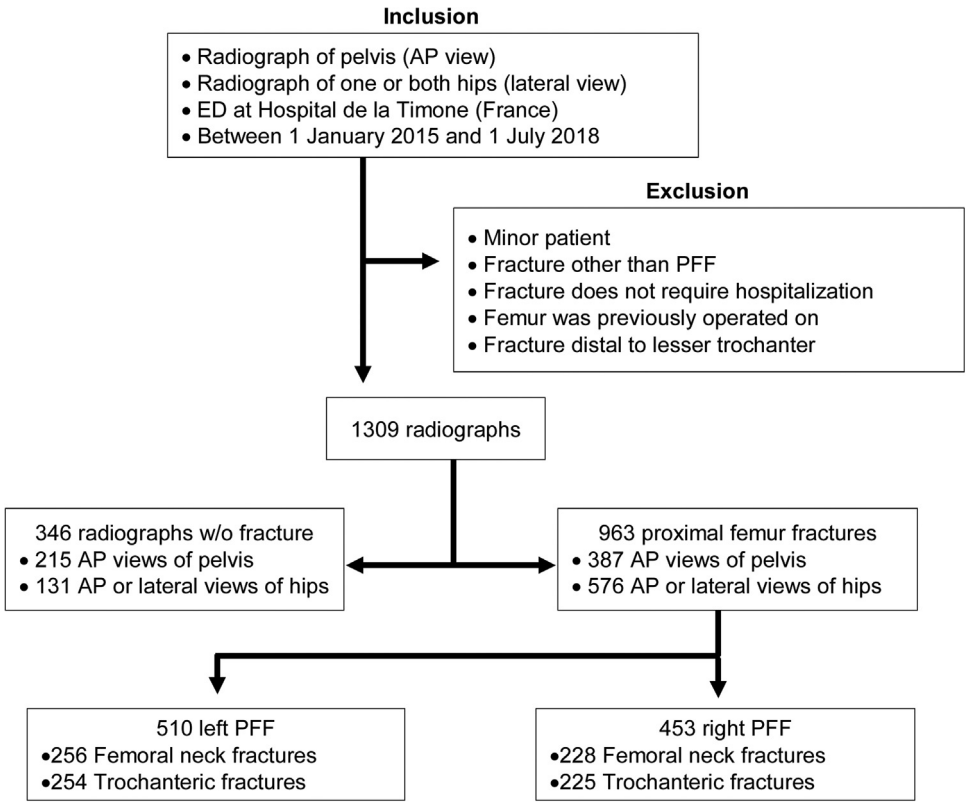

Fig. 1. Flow chart (AP: anteroposterior, ED: emergency department, PFF: proximal femoral fracture, w/o: without).

the performance of a trained radiologist. AI may turn out to be most beneficial in assisting non-radiologists to make a diagnosis.

In the emergency department (ED), 3.7% to 4.1% of fractures are missed [21,22], and 70% to 80% of missed diagnoses in the ED are fractures [23,24]. These errors are more likely at night [21], because of overwork and because no orthopedic specialists are available. The consequences for the patient can be serious: loss of function and autonomy, increased risk of nonunion, post-traumatic osteoarthritis [25]. The counter this deficit, programs to assist in the radiographic diagnosis of proximal femur fractures (PFF) [26,27], distal radius fractures [25] and general trauma [28] have been developed. In some cases, their accuracy surpasses that of experienced orthopedic surgeons [27]. The diagnostic error rate of ED physicians is reduced [25]. Since PFF are a public health problem, a more accurate and refined diagnosis can optimize the treatment, reduce post-traumatic morbidity and improve clinical recovery [29,30]. Thus, an effective, low-cost software that aids radiographic diagnoses would be a relevant tool for current clinical practice, although this concept has rarely been evaluated in orthopedics for the proximal femur. This led us to conduct a prospective study with the goals of:

- programming deep learning software to help make the diagnosis of PFF on radiographs;
- to evaluate its performance.

We hypothesized that deep learning software can be programmed to help make the diagnosis of PFF using a limited number of radiographs.

2. Materials and methods

2.1. Database

After approval from our facility's Data Protection Committee, 1309 radiographs from 623 patients were collected retrospectively from an imaging database of patients who had been hospitalized at our facility. The inclusion criteria were the availability of a high-quality AP radiograph of the pelvis or AP and lateral views of one or both hips, in a patient who presented at the ED between 1 January 2015 and 1 July 2018. Both AP and lateral views were used to match the realities of clinical practice; these views are taken routinely when a patient presents with this type of injury. This allows the software to be trained to read different radiographic views, without reducing the database size. Excluded were patients who were less than 18 years of age, who had a fracture other than in the proximal femur, who previously had surgery on the femur, who had a fracture that did not require hospitalization, or who had a fracture distal to the trochanteric area. All the radiographs were anonymized. These data are summarized in the flow chart in Fig. 1.

All the radiographs were made on the Digital Diagnost FLEXR^{TM} (Philips, Amsterdam, The Netherlands) using the following parameters: Study Time 155919.174000, Series Time 155920.000000, Acquisition Time 155921, Content Time 155921.

2.2. Annotation of radiographs

All the radiographic images were processed in JPEG format and were annotated using the Rect Label^{TM} software (https://rectlabel.com) by an experienced orthopedic and trauma surgeon (MO–MD, PhD). Two groups were made: a healthy group, corresponding to healthy landmarks and a fracture group, corresponding to fractured areas.

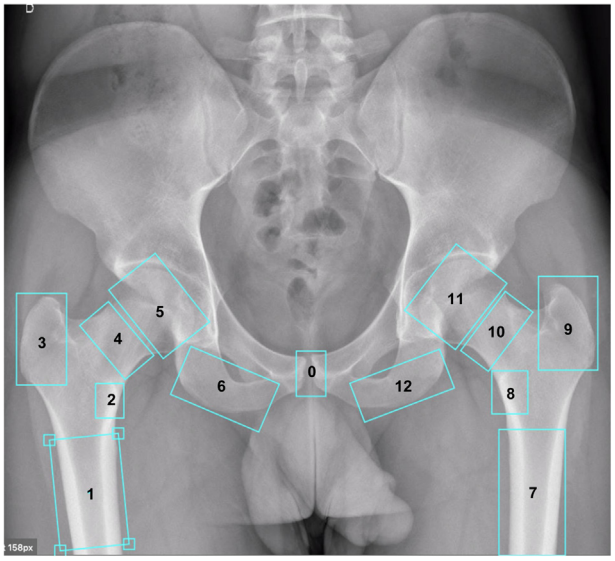

Fig. 2. Landmarks on radiographs of healthy hips.

First, the healthy landmarks were identified to allow our program to differentiate a fracture-free zone from a fractured one. Nineteen landmarks were chosen: 13 bone landmarks and 6 metal landmarks. Certain patients had metal implants in their contralateral femur or lumbar spine. The latter were identified to allow the software to recognized them as "healthy". However, if the fracture affected the side that had previously been operated on, these patients were excluded from the analysis.

Among the bone landmarks, 12 were lateralized as "right" or "left", with the pubic symphysis acting as the central landmark. The metal landmarks were not lateralized to simplify the learning phase. These healthy landmarks are shown in Fig. 2.

The 963 fractures were annotated this way. The fractures were classified based on the AO classification for PFF [31]. Despite debatable intra- and inter-rater reliability [32], the association of a treatment decision tree to each subgroup makes this classification relevant for use by a non-specialized physician. However, given the small number of radiographs annotated in each AO subgroup, a satisfactory success rate could not be achieved for classifying the fractures. Only our algorithm's ability to distinguish between femoral neck fractures and trochanteric fractures was evaluated during the test phase.

To simplify our program's learning, the fracture annotation did not specify which side it was on. The 31A1.1 subgroup represents isolated greater trochanter or lesser trochanter fractures. Since these fractures do not require hospitalization, none were found in our study. All the data are given in Table 1.

2.3. Increasing the database size

All the images were turned 180°, inverting left and right, to double the sample size. Next, the images were rotated, and their width, height and magnification modified randomly by 0 to 15% of their original value. These manipulations increased 8-fold the total number of radiographs, resulting in 10,472 analyzable images. This technique has previously been described and validated [33].

2.4. Programming

The deep learning algorithm was programmed on Tensorflow™ software (Google Brain, Santa Clara, Ca, USA, tensorflow.org). This is the automatic learning tool that is currently used the most in the AI domain. The Lobe neuronal network (Microsoft, Redmond, Washington, USA, lobe.ai), consisting of multiple convolutional layers, made up the architecture for our algorithm. This network has not previously been trained to read medical images.

2.5. Training phase

Among the 10,472 images obtained after the database expansion, 90% (9425 radiographs) were used for the training phase. The remaining 1074 radiographs were reserved for the test phase. The training phase was divided into two parts: a learning phase with 8351 radiographs (80% of the total) and a validation phase with 1074 (10% of the total). A statistically similar proportion of pathological radiographs were found in each phase, corresponding to the ratio in the entire sample: 25% to 30% healthy radiographs versus 70% to 75% pathological radiographs. Our algorithm's performance was evaluated by analyzing its success rate for detecting fractures during the test phase.

2.6. Statistical analysis

The primary aim of this study was to evaluate diagnostic performance. The data were described using the senior author's analysis for each radiograph (reference). From this comparison, the true and false positive/true and false negatives were calculated along with the positive predictive value (PPV), negative predictive value (NPV), sensitivity and specificity of the tests.

3. Results

One thousand seventy-four radiographs were submitted to our trained algorithm. Its ability to recognize healthy landmarks and detect fractures was evaluated.

3.1. Facture detection

The fracture detection results are provided in Table 2.

Table 1
AO classification of annotated fractures [31].

Trochanteric fractures $n = 479$		Femoral neck fractures $n = 484$	
31A1.1	$n = 0$	31B1.1	$n = 26$
31A1.2	$n = 40$	31B1.2	$n = 12$
31A1.3	$n = 105$	31B1.3	$n = 20$
31A2.1	$n = 244$	31B2.1	$n = 204$
31A2.3	$n = 42$	31B2.2	$n = 195$
31A3.1	$n = 10$	31B2.3	$n = 8$
31A3.2	$n = 7$	31B3	$n = 19$
31A3.3	$n = 31$		

Table 2
Fracture detection results.

	n	Augmented *n*	Tested *n*	Identified *n*	% Identified
Left femoral neck fracture	256	2048	198	122	61.61%
Left trochanteric fracture	254	2032	208	141	67.79%
Right femoral neck fracture	228	1824	193	116	60.10%
Right trochanteric fracture	225	1800	175	115	65.71%

Femoral neck fractures

	Patients with fracture	Patients without fracture
Positive test	238	213
Negative test	153	443
	Se 60.87%	VPP 55.22%
	Sp 67.53%	VPN 74.33%

TrochTrochanteric fractures

	Patients with fracture	Patients without fracture
Positive test	256	202
Negative test	127	462
	Se 66.84%	VPP 55.89%
	Sp 69.58%	VPN 78.44%

Se: Sensitivity/Sp: Specificity/PPV: Positive predictive value, NPV: Negative predictive value.

3.2. Femoral neck fractures

In all, 391 radiographs with femoral neck fractures were tested. The algorithm identified 238 fractures–122 left (62%, 122/198) and 116 right (60%, 116/193)–thus a success rate of 61% (238/391) (Table 2). With 213 false positives versus 238 true fractures identified, our diagnostic tool had a PPV of 55%. There were fewer false negatives ($n = 153$) relative to the number of radiographs correctly identified as not having a femoral neck fracture ($n = 443$), which means our diagnostic tool had a NPV of 74% (Table 2). The sensitivity was 61% and the specificity was 67% for detecting femoral neck fractures (Table 2).

3.3. Trochanteric fractures

In all, 383 radiographs with trochanteric fractures were tested. The algorithm successfully identified 256 fractures–141 on the left side (68%, 141/208) and 115 on the right side (66%, 115/175)–thus an overall success rate of 67% (256/383) (Table 2). There were 202 false positives. Given that 256 trochanteric fractures were correctly identified, the PPV was 56%. There were 443 false negatives. Since 153 fractures were missed, the PPV of our tool was 78% (Table 2). The sensitivity was 67% and the specificity was 70% for detecting trochanteric fractures (Table 2).

3.4. Healthy radiographic markers

On average, 60% of the healthy radiographic markers were correctly identified by our algorithm. All the results are given in Table 3. The highest success rate was found for the pubic symphysis (100%), femoral diaphysis (84% on the right, 478/572, and 82% on the left, 492/603) and greater trochanters (75% on the right, 352/472, and 86% on the left, 372/431). The other landmarks were identified in only about half the cases. While many ischiums were used during the training phase, they were located successfully in only 60% (423/709) and 62% (442/715) of cases for the right and left sides, respectively. The most common error was incorrect identification of the side.

The metal landmarks were the most difficult to identify, with a success rate of only 47%. While there were few lumbar fusion constructs in the database, their central and easily identifiable nature resulted in good performance of our algorithm.

Table 3
Landmarks on radiographs of healthy hips: results.

	n	Augmented *n*	Tested *n*	Identified *n*	% Identified
Right femoral shaft	726	5808	572	478	83.57%
Right lesser trochanter	660	5280	513	234	45.61%
Right greater trochanter	575	4600	472	352	74.58%
Right femoral neck	276	2208	213	108	50.70%
Right femoral head	603	4824	493	254	51.52%
Right ischium	895	7160	709	423	59.66%
Left femoral shaft	741	5928	603	492	81.59%
Left lesser trochanter	676	5408	532	274	51.50%
Left greater trochanter	534	4272	431	372	86.31%
Left femoral neck	267	2136	216	102	47.22%
Left femoral head	573	4584	449	273	60.80%
Left ischium	906	7248	715	442	61.81%
Pubic symphysis	924	7392	742	742	100%
Gamma nail	29	232	24	10	41.67%
Total hip arthroplasty	32	256	22	7	31.82%
Hemiarthroplasty	26	208	23	12	52.17%
Dynamic Hip Screw	4	32	2	1	50%
Lumbar instrumentation	9	72	8	6	75%
Screw fixation of femoral neck	6	48	3	1	33.33%

4. Discussion

To be relevant in clinical practice, our algorithm needs to have equal or better diagnostic accuracy than an experienced ED physician. In the literature, it is said that an ED physician without specific trauma surgery training has a minimum sensitivity of about 80% [25] for detecting fractures, versus nearly 90% for an experienced orthopedic surgeon [27], with a specificity in both cases greater than 10 points. With a maximum sensitivity of 67% and specificity of 70%, our algorithm was not good enough, thus our hypothesis is rejected.

These results can be explained by the small number of images used to train our program, despite the 8-fold increase of our 1309 radiographs. Programs described in the literature that achieve diagnostic accuracy of 90% were developed with a much larger database of images. Lindsey et al. [25] trained their deep learning program to detect wrist fractures using more than 135,000 radiographs annotated by eight experienced orthopedic surgeons. This explains the excellent results with that program, allowing ED physicians to increase their diagnostic sensitivity and specificity by 10 points to above 90%, with mean reduction of 47% in the error rate.

When Olczak et al. [28] published the first article on a deep learning program for fracture detection in 2017, they used a database of 256,000 wrist, hand and ankle radiographs. Despite this, the diagnostic accuracy for detecting a fracture was only 83%. Deep learning had only been used very recently in the medical field when this program was developed; the lack of experience may explain these disappointing results.

In April 2019, Cheng et al. [26] published the results of an algorithm trained for diagnosing proximal femur fractures. This program, which resembles ours the most, was trained using nearly 30,000 AP radiographs of the pelvis. This large number of images and use of a single view increased the reproducibility and explains the high diagnostic accuracy of this tool: 98% sensitivity with only 2% false negatives.

Our program, which was designed specifically to detect fractures, was not trained to recognize non-medical images. There are published examples of deep learning algorithms, trained to recognized non-medical images, that were effectively converted to radiographic diagnosis using "transfer learning". In these cases, a small number of radiographs were used: slightly more than 1000 for Adams et al. [34] and 11,000 for Kim and MacKinnon [33]. Despite this, the goal of 90% accuracy was achieved. However, hundreds of thousands non-medical images were used for the programming beforehand. These good results need to be put into perspective when we attempt to determine the true impact of AI on daily radiology practice. For example, in February 2019, Adams et al. [34] showed that, with only 1 hour of practice, a person without medical training could rival an AI trained by transfer learning for the diagnosis of femoral neck fractures.

Thus our program is not competitive, given the performance of other deep learning algorithms described in the literature. Relative to all the previously mentioned AI studies, a smaller number of radiographs were used during its programming. Despite this fact, its NPV was nearly 80% and a maximum sensitivity and specific of nearly 70%. These results are encouraging. Since we used 25 times fewer images than Olczak et al. [28], we can presume that adding new annotated radiographs would greatly improves its performance.

Also, the AO classification of fractures would allow a decision tree to be attributed to each subgroup for their treatment. While the smaller number of available radiographs did not allow us to test our algorithm on this feature, this is an interesting avenue in the future when a larger database will be available.

We can now ask questions about the future of AI in the medical world. The mass of data available online can now be used by deep learning algorithms. Ausiello and Shaw [35] content that "Quantitative Human Phenotyping" is the next great step in medical research. They point to the benefits that we could gain from exploring public personal data available online in multiple domains, such as sociology and medicine. Common use of DNA sequencing databases on the Internet could help to accelerate genetics research. Esteva et al. [11] advanced the hypothesis that a database of several million photographs of dermatological lesions will be available within a few years thanks to smartphones, and could help to optimize the diagnostic performance of their algorithms. Use of AI in medical practice, which is still marginal, could very likely become more widespread in the coming years given the pace of technical advances.

Our study had several limitations: image labelling was done by only one orthopedic surgeon, the ROC curve and its area under the curve were not calculated, and transfer learning was not used to optimize the performance of our software. While using a single experienced orthopedic surgeon for annotation is certainly a methodological flaw, it does not directly impact our software's performance. Furthermore, the sensitivity, specificity, PPV and NPV are reliable, relevant parameters that could be reinforced by the ROC, although these parameters can be interpreted in their current state. Lastly, not having recourse to transfer learning explains the lack of power of our software and reinforces our conclusion: effective software to help with radiographic diagnoses can only be programmed using several thousands of radiographs or with the help of transfer learning.

5. Conclusion

The performance of our program was not good enough to be clinically relevant. Its main flaw was the small number of radiographs available to us, which did not provide an optimal training phase. Programming of a deep learning software with sufficient diagnostic accuracy can only be done with several tens of thousands of radiographs, or by using transfer learning.

Disclosure of interest

Sylvain Guy, Christophe Jacquet and Damien Tsenkoff have no conflicts of interest to declare. Jean-Noël Argenson is a consultant for Zimmer. Matthieu Ollivier is a consultant for Newclip Technics, Arthrex and Smith & Nephew.

Funding

No financing was received for this study.

Authors' contributions

Sylvain Guy: Conception and design of the study, acquisition of data, experimentation, analysis and interpretation of data, statistics.

Christophe Jacquet: Revising the article critically for important intellectual content.

Damien Tsenkoff: Data scientist.

Jean-Noël Argenson: Revising the article critically for important intellectual content.

Matthieu Ollivier: Experimentation, revising the article critically for important intellectual content, final approval of the version to be submitted.

References

[1] Xiong W, Droppo J, Huang X, Seide F, Seltzer ML, Stolcke A, et al. Toward Human Parity in Conversational Speech Recognition. IEEEACM Trans Audio Speech Lang Proc 2017;25:2410–23, http://dx.doi.org/10.1109/TASLP.2017.2756440.

[2] Pendleton SD, Andersen H, Du X, Shen X, Meghjani M, Eng YH, et al. Perception, Planning, Control, and Coordination for Autonomous Vehicles. Machines 2017;5:6, http://dx.doi.org/10.3390/machines5010006.
[3] Mnih V, Kavukcuoglu K, Silver D, Rusu AA, Veness J, Bellemare MG, et al. Human-level control through deep reinforcement learning. Nature 2015;518:529–33, http://dx.doi.org/10.1038/nature14236.
[4] Moravčík M, Schmid M, Burch N, Lisý V, Morrill D, Bard N, et al. DeepStack: Expert-level artificial intelligence in heads-up no-limit poker. Science 2017;356:508–13, http://dx.doi.org/10.1126/science.aam6960.
[5] Silver D, Huang A, Maddison CJ, Guez A, Sifre L, van den Driessche G, et al. Mastering the game of Go with deep neural networks and tree search. Nature 2016;529:484–9, http://dx.doi.org/10.1038/nature16961.
[6] LeCun Y, Bengio Y, Hinton G. Deep learning. Nature 2015;521:436–44, http://dx.doi.org/10.1038/nature14539.
[7] Deo RC. Machine Learning in Medicine. Circulation 2015;132:1920–30, http://dx.doi.org/10.1161/CIRCULATIONAHA.115.001593.
[8] Alipanahi B, Delong A, Weirauch MT, Frey BJ. Predicting the sequence specificities of DNA- and RNA-binding proteins by deep learning. Nat Biotechnol 2015;33:831–8, http://dx.doi.org/10.1038/nbt.3300.
[9] Yuan Y, Shi Y, Li C, Kim J, Cai W, Han Z, et al. DeepGene: an advanced cancer type classifier based on deep learning and somatic point mutations. BMC Bioinformatics 2016;17:476, http://dx.doi.org/10.1186/s12859-016-1334-9.
[10] Gulshan V, Peng L, Coram M, Stumpe MC, Wu D, Narayanaswamy A, et al. Development and Validation of a Deep Learning Algorithm for Detection of Diabetic Retinopathy in Retinal Fundus Photographs. JAMA 2016;316:2402–10, http://dx.doi.org/10.1001/jama.2016.17216.
[11] Esteva A, Kuprel B, Novoa RA, Ko J, Swetter SM, Blau HM, et al. Dermatologist-level classification of skin cancer with deep neural networks. Nature 2017;542:115–8, http://dx.doi.org/10.1038/nature21056.
[12] Cireşan DC, Giusti A, Gambardella LM, Schmidhuber J. Mitosis detection in breast cancer histology images with deep neural networks. Med Image Comput Comput-Assist Interv MICCAI Int Conf Med Image Comput Comput-Assist Interv 2013;16:411–8.
[13] Sirinukunwattana K, Ahmed Raza SE, Yee-Wah Tsang null, Snead DRJ, Cree IA, Rajpoot NM. Locality Sensitive Deep Learning for Detection and Classification of Nuclei in Routine Colon Cancer Histology Images. IEEE Trans Med Imaging 2016;35:1196–206, http://dx.doi.org/10.1109/TMI.2016.2525803.
[14] Litjens G, Kooi T, Bejnordi BE, Setio AAA, Ciompi F, Ghafoorian M, et al. A survey on deep learning in medical image analysis. Med Image Anal 2017;42:60–88, http://dx.doi.org/10.1016/j.media.2017.07.005.
[15] Hosny A, Parmar C, Quackenbush J, Schwartz LH, Aerts HJWL. Artificial intelligence in radiology. Nat Rev Cancer 2018;18:500–10, http://dx.doi.org/10.1038/s41568-018-0016-5.
[16] Shen D, Wu G, Suk H-I. Deep Learning in Medical Image Analysis. Annu Rev Biomed Eng 2017;19:221–48, http://dx.doi.org/10.1146/annurev-bioeng-071516-044442.
[17] Erickson BJ, Korfiatis P, Akkus Z, Kline TL. Machine Learning for Medical Imaging. Radiogr Rev Publ Radiol Soc N Am Inc 2017;37:505–15, http://dx.doi.org/10.1148/rg.2017160130.
[18] Lehman CD, Wellman RD, Buist DSM, Kerlikowske K, Tosteson ANA, Miglioretti DL, et al. Diagnostic Accuracy of Digital Screening Mammography With and Without Computer-Aided Detection. JAMA Intern Med 2015;175:1828–37, http://dx.doi.org/10.1001/jamainternmed.2015.5231.
[19] Kooi T, Litjens G, van Ginneken B, Gubern-Mérida A, Sánchez CI, Mann R, et al. Large scale deep learning for computer aided detection of mammographic lesions. Med Image Anal 2017;35:303–12, http://dx.doi.org/10.1016/j.media.2016.07.007.
[20] Rajkomar A, Lingam S, Taylor AG, Blum M, Mongan J. High-Throughput Classification of Radiographs Using Deep Convolutional Neural Networks. J Digit Imaging 2017;30:95–101, http://dx.doi.org/10.1007/s10278-016-9914-9.
[21] Hallas P, Ellingsen T. Errors in fracture diagnoses in the emergency department–characteristics of patients and diurnal variation. BMC Emerg Med 2006;6:4, http://dx.doi.org/10.1186/1471-227X-6-4.
[22] Wei C-J, Tsai W-C, Tiu C-M, Wu H-T, Chiou H-J, Chang C-Y. Systematic analysis of missed extremity fractures in emergency radiology. Acta Radiol Stockh Swed 1987;47(2006):710–7, http://dx.doi.org/10.1080/02841850600806340.
[23] Guly HR. Diagnostic errors in an accident and emergency department. Emerg Med J EMJ 2001;18:263–9, http://dx.doi.org/10.1136/emj.18.4.263.
[24] Leeper WR, Leeper TJ, Vogt KN, Charyk-Stewart T, Gray DK, Parry NG. The role of trauma team leaders in missed injuries: does specialty matter? J Trauma Acute Care Surg 2013;75:387–90, http://dx.doi.org/10.1097/TA.0b013e31829fa32.
[25] Lindsey R, Daluiski A, Chopra S, Lachapelle A, Mozer M, Sicular S, et al. Deep neural network improves fracture detection by clinicians. Proc Natl Acad Sci U S A 2018;115:11591–6, http://dx.doi.org/10.1073/pnas.1806905115.
[26] Cheng C-T, Ho T-Y, Lee T-Y, Chang C-C, Chou C-C, Chen C-C, et al. Application of a deep learning algorithm for detection and visualization of hip fractures on plain pelvic radiographs. Eur Radiol 2019, http://dx.doi.org/10.1007/s00330-019-06167-y.
[27] Urakawa T, Tanaka Y, Goto S, Matsuzawa H, Watanabe K, Endo N. Detecting intertrochanteric hip fractures with orthopedist-level accuracy using a deep convolutional neural network. Skeletal Radiol 2019;48:239–44, http://dx.doi.org/10.1007/s00256-018-3016-3.
[28] Olczak J, Fahlberg N, Maki A, Razavian AS, Jilert A, Stark A, et al. Artificial intelligence for analyzing orthopedic trauma radiographs. Acta Orthop 2017;88:581–6, http://dx.doi.org/10.1080/17453674.2017.1344459.
[29] Shin WC, Moon NH, Jang JH, Jeong JY, Suh KT. Three-dimensional analyses to predict surgical outcomes in non-displaced or valgus impaction fractures of the femoral neck: A multicenter retrospective study. Orthop Traumatol Surg Res 2019;105:991–8, http://dx.doi.org/10.1016/j.otsr.2019.03.016.
[30] Oba T, Makita H, Inaba Y, Yamana H, Saito T. New scoring system at admission to predict walking ability at discharge for patients with hip fracture. Orthop Traumatol Surg Res 2018;104:1189–92, http://dx.doi.org/10.1016/j.otsr.2018.07.024.
[31] Meinberg E, Agel J, Roberts C, Karam M, Kellam J. Fracture and Dislocation Classification Compendium–2018. J Orthop Trauma 2018;32:S1–10, http://dx.doi.org/10.1097/BOT.0000000000001063.
[32] Masionis P, Uvarovas V, Mazarevičius G, Popov K, Venckus Š, Baužys K, et al. The reliability of a Garden, AO and simple II stage classifications for intracapsular hip fractures. Orthop Traumatol Surg Res 2019;105:29–33, http://dx.doi.org/10.1016/j.otsr.2018.11.007.
[33] Kim DH, MacKinnon T. Artificial intelligence in fracture detection: transfer learning from deep convolutional neural networks. Clin Radiol 2018;73:439–45, http://dx.doi.org/10.1016/j.crad.2017.11.015.
[34] Adams M, Chen W, Holcdorf D, McCusker MW, Howe PD, Gaillard F. Computer vs human: Deep learning versus perceptual training for the detection of neck of femur fractures. J Med Imaging Radiat Oncol 2019;63:27–32, http://dx.doi.org/10.1111/1754-9485.12828.
[35] Ausiello D, Shaw S. Quantitative Human Phenotyping: The Next Frontier in Medicine. Trans Am Clin Climatol Assoc 2014;125:219–28.