Osteoporosis International (2021) 32:209–212
https://doi.org/10.1007/s00198-020-05737-x

EDITORIAL

Evolution in fracture risk assessment: artificial versus augmented intelligence

D. Hans[1] · E. Shevroja[1] · W. D. Leslie[2]

Received: 3 November 2020 / Accepted: 8 November 2020 / Published online: 7 January 2021

> "Nothing in life is to be feared, it is only to be understood. Now is the time to understand more, so that we may fear less." (Marie Curie)

Artificial intelligence in medicine comes with enormous promise but also potential pitfalls. Liu Y et al. [1] highlight the need for a cautious and critical approach to evaluate machine learning tools, as with any diagnostic tool, that must be supported by clinical judgment: "…clinical gestalt plays a crucial role in evaluating whether the results are believable. Results that substantially exceed what even such a hypothetical expert is capable of should be scrutinized and validated carefully."

The transition from the physician's handwritten notes to electronic health records and a plethora of digital data ushered in the era of Big Data in medicine. Classical hypothesis-driven research is giving way to data-driven research, with opportunities to pursue novel questions and directions raised from the data itself. The statistical approaches currently used to explore this expanding data universe are often drawn from the field of artificial intelligence. The principle of AI is to mimic the thinking and decision-making capabilities of humans using a variety of algorithmic tools.

Clinical decision-making, as much as art as science, is the final outcome of a complex process that rest on scientific knowledge and clinical experience gained through years of training and practice. Evidence-based medicine (EBM) and clinical trials represent the pinnacle of scientific decision-making. The ability to exploit Big Data with AI offers the potential to greatly accelerate the experience-based component of the decision-making process. This interplay of EBM and AI can ultimately enhance the physician's performance.

Within the broad discipline of AI, the subfields of machine learning (ML) and deep learning (DL) are currently of greatest relevance to medical practice (Fig. 1) [1–4].

The two main approaches to ML are supervised or unsupervised learnings [1, 2]. Supervised learning uses a given set of input features and one or more outcomes (labels) as the basis for model training. The model is iteratively trained to minimize prediction error when comparing samples drawn from the data with a target reference standard, also called ground truth. Supervised ML for predicting a known outcome is the most widely used approach at present. Unsupervised learning does not use any labeling information and aims to group data by shared properties. This helps to discover structure in the data, such as identifying clusters of patients at similar risk or selecting variables most strongly correlated with an outcome. DL, a specific group of ML methods, uses multi-layered arithmetic operations (sometimes hundreds of layers containing many millions of individual calculations) in order to model the complex non-linear relationships between data inputs and outputs.

While AI/ML is designed to output a simple answer, the underlying process to get there is extremely complex and requires attention to numerous technical details. This is analogous to the diagnostic process in medicine. The final diagnosis conceals a non-linear reasoning pathway that incorporates medical knowledge and experience with clinical clues from the history, physical examination, and investigations. The classical pipeline for developing and implementing a supervised ML model is based on the subsequent steps shown in Fig. 2.

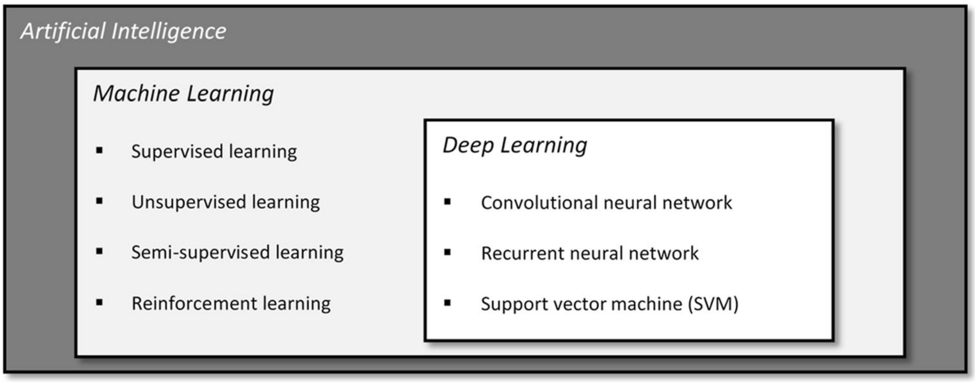
Although each of these steps is important, data preparation deserves special attention given the data-driven nature of ML. Garbage-in-garbage-out (GIGO) describes the importance of data quality. Access to Big Data is not enough—we must ensure High Quality Big Data. Failure to adhere to this

✉ D. Hans
didier.hans@ascendys.ch

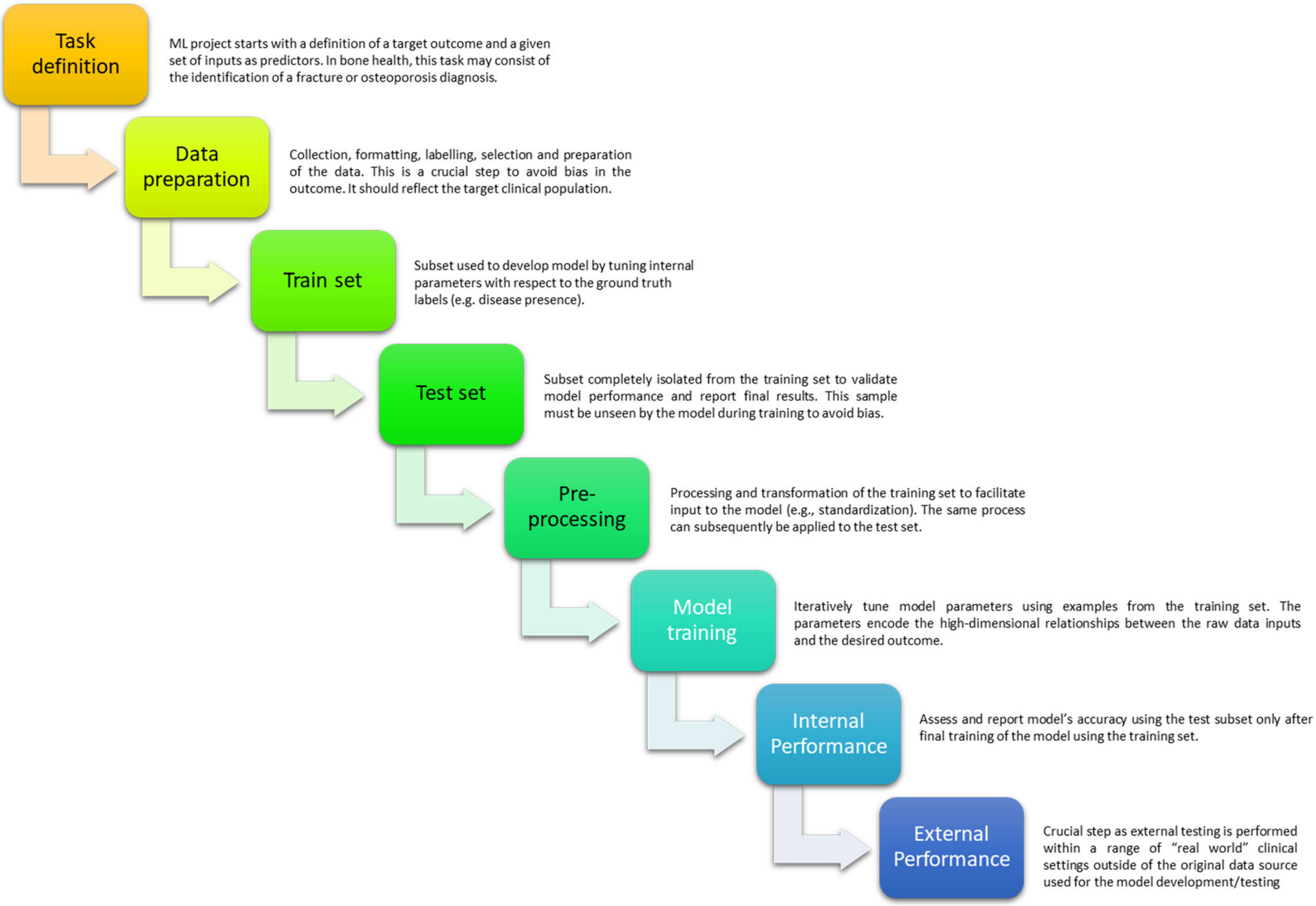
[1] Interdisciplinary Center of Bone Diseases, Bone and Joint Department, Lausanne University Hospital and Lausanne University, Lausanne, Switzerland

[2] Department of Medicine, University of Manitoba, Winnipeg, Manitoba, Canada

Fig. 1 Hierarchical classification with examples of artificial intelligence, machine learning, and deep learning



principle can lead to biased or even erroneous results. Companies are rushing to provide off-the-shelf platforms using pre-defined algorithms for democratizing AI access. In theory, one only needs to load of the data and specify a few parameters, voilà—a fully trained convolutional neural network (CNN)! However, caveat emptor. Any biases in the data collection or labeling (e.g., establishing the ground truth) would automatically generate systematic errors in the predictions that machines would now perform repeatedly. In contrast to carefully collected and adjudicated research data, Big Data comes from "real-world" sources, which are comparatively "dirty." Nothing is free, and the cost of data quantity is questionable quality, which can affect the reliability of the derived ML products. In summary, any deviation from the eight steps described previously can lead to overly optimistic (or more rarely pessimistic) results, thereby threatening clinical reliability of the results. Accordingly, it is important not to underreport model details and clinical information as any lack of reporting transparency impedes effective comparisons, model reproducibility, and clinical use [4, 5].



ML project starts with a definition of a target outcome and a given set of inputs as predictors. In bone health, this task may consist of the identification of a fracture or osteoporosis diagnosis.

Collection, formatting, labelling, selection and preparation of the data. This is a crucial step to avoid bias in the outcome. It should reflect the target clinical population.

Subset used to develop model by tuning internal parameters with respect to the ground truth labels (e.g. disease presence).

Subset completely isolated from the training set to validate model performance and report final results. This sample must be unseen by the model during training to avoid bias.

Processing and transformation of the training set to facilitate input to the model (e.g., standardization). The same process can subsequently be applied to the test set.

Iteratively tune model parameters using examples from the training set. The parameters encode the high-dimensional relationships between the raw data inputs and the desired outcome.

Assess and report model's accuracy using the test subset only after final training of the model using the training set.

Crucial step as external testing is performed within a range of "real world" clinical settings outside of the original data source used for the model development/testing

Fig. 2 Simplified model development flowchart for supervised learning

Many of the early successes of AI in medicine have been in image-intensive specialties, such as radiology, pathology, ophthalmology, and cardiology [6]. Clinical risk prediction, diagnostics, and therapeutics are more challenging. Hence, AI is still relatively novel in the osteoporosis field. A query on PubMed indicates an exponential increase in AI publications since 2010 with more than 38,000 articles, with over 10,000 in the last year alone. In contrast, fewer than 100 of these were in the field of osteoporosis, although this is following the same exponential trajectory with the majority of studies published during the last 2–3 years. Efforts have been made in osteoporosis diagnosis and classification, bone mineral density assessment, fracture detection, fracture risk estimation, and bone image segmentation [7–14]. The majority of these articles used opportunistic data—particularly in imaging.

Accurate fracture risk estimation is crucial in osteoporosis management and the first step in bone health clinical evaluation. Widely used fracture risk assessment tools (e.g., FRAX®) are based on classical statistical approaches informed by clinical expertise in osteoporosis [15]. For instance, FRAX was developed and validated in various large population-based datasets. Each individual clinical risk factor (twelve in total) was incorporated into FRAX based on a solid scientific rationale and a supporting meta-analysis. Anticipating the rise of Big Data, FRAX is an example of how evidence-based hypotheses drive data analysis and find their way into clinical utility.

In this issue of Osteoporosis International, De Vries et al. [16] compared fracture risk prediction from classical techniques (Cox regression) with AI-/ML-based survival models (random survival forests, RSF, and artificial neural network, ANN-DeepSurv). Their study was conducted in a sample of 7578 post-fracture individuals, relatively large by conventional measures though not by current ML standards. Their data-driven hypothesis-free investigation aimed to compare the performance of the models and to identify, if possible, novel risk factors. This study reminds us about the wealth of electronic data that is increasingly available to researchers from electronic health record databases. Although FRAX performs well in clinical practice, it still ignores large amounts of potentially valuable patient information. The use of these additional data sources, such as in De Vries et al., can suggest novel hypotheses, risk factors, and disease/health determinants. Despite examining more than 40 clinical and laboratory variables in their predictive models, and contrary to expectation, Cox regression outperformed the AI/ML models. In part, this may reflect the use of more sophisticated approaches with the Cox regression (LASSO variable selection, non-linear transforms including restricted cubic splines) and rather simplistic ML architectures (only 2 layers in ANN-DeepSurv). One would anticipate that ML performance would improve with more complex architectures and sufficient data to avoid overfitting. The identification of overlapping predictors (e.g., age, hip T-score, time since menopause) provides face validity that the approaches are responsive to similar signals in the data. ML was also able to identify plausible risk factors that were omitted from the Cox model (e.g., vertebral fracture, lumbar spine T-score) and to propose some novel risk factors (e.g., plasma albumin, breastfeeding), both worthy of future study.

We are optimistic regarding in the future of medical AI and for the osteoporosis field in particular. In 2018, the Food and Drug Administration (FDA) approved the OsteoDetect (Imagen), an AI software based upon DL to identify and highlight distal radius fractures during the review of posterior-anterior and lateral radiographs of adult wrists as an adjunct to the clinician's review and clinical judgment [17]. Thus, AI has already found its way to addressing important tasks in the overall osteoporosis clinical management. Nevertheless, healthy skepticism should balance zeal to see this science move forward and temper our enthusiasm to see AI integrated into clinical applications. A recent systematic review of medical imaging DL algorithms between 2010 and June 2019 found that almost all were retrospective, non-randomized, at high risk of bias, and deviated from existing reporting standards [18]. Moreover, data and code availability were lacking in most studies, yet only a minority stated that further prospective studies or trials were required. Aside from technical issues related to model reliability and reporting transparency, AI/ML raises prickly new questions that have yet to be answered: Who owns the data used to initially train the algorithm and what are the rights of the patient to control their personal information? How are individual privacy concerns balanced against making the dataset (which can be highly detailed in "real-world" data) available for independent validation? Once approved, how can one provide "stewardship" over changes to the AI/ML algorithm (mostly cloud based) without stifling its unique ability to evolve and improve?

To conclude, we believe healthcare will see increasing synergy between human and artificial intelligences, where the latter will enhance a physician's performance and support well-informed clinical decision-making—namely augmented intelligence. AI should be seen as yet another tool for improving the quality of patient care. Not to be feared but to be understood, as we explore AI's unique strengths and limitations.

Compliance with ethical standards

Conflict of interest Didier Hans, Enisa Shevroja, and William D Leslie declare that they have no conflict of interest related to this editorial.

References

1. Liu Y, Chen PC, Krause J, Peng L (2019) How to read articles that use machine learning: users' guides to the medical literature. Jama 322(18):1806–1816. https://doi.org/10.1001/jama.2019.16489
2. Rajkomar A, Dean J, Kohane I (2019) Machine learning in medicine. N Engl J Med 380(14):1347–1358. https://doi.org/10.1056/NEJMra1814259
3. Beam AL, Kohane IS (2018) Big data and machine learning in health care. Jama 319(13):1317–1318. https://doi.org/10.1001/jama.2017.18391
4. LeCun Y, Bengio Y, Hinton G (2015) Deep learning. Nature 521(7553):436–444. https://doi.org/10.1038/nature14539
5. Liu X, Faes L, Kale AU, Wagner SK, Fu DJ, Bruynseels A, Mahendiran T, Moraes G, Shamdas M, Kern C, Ledsam JR, Schmid MK, Balaskas K, Topol EJ, Bachmann LM, Keane PA, Denniston AK (2019) A comparison of deep learning performance against health-care professionals in detecting diseases from medical imaging: a systematic review and meta-analysis. Lancet Digital Health 1(6):e271–e297. https://doi.org/10.1016/S2589-7500(19)30123-2
6. Maddox TM, Rumsfeld JS, Payne PRO (2019) Questions for artificial intelligence in health care. Jama 321(1):31–32. https://doi.org/10.1001/jama.2018.18932
7. Nazia Fathima SM, Tamilselvi R, Parisa Beham M, Sabarinathan D (2020) Diagnosis of osteoporosis using modified U-net architecture with attention unit in DEXA and X-ray images. J X-ray Sci Technol 28(5):953–973. https://doi.org/10.3233/xst-200692
8. Pan Y, Shi D, Wang H, Chen T, Cui D, Cheng X, Lu Y (2020) Automatic opportunistic osteoporosis screening using low-dose chest computed tomography scans obtained for lung cancer screening. Eur Radiol 30(7):4107–4116. https://doi.org/10.1007/s00330-020-06679-y
9. Nam KH, Seo I, Kim DH, Lee JI, Choi BK, Han IH (2019) Machine learning model to predict osteoporotic spine with Hounsfield units on lumbar computed tomography. J Korean Neurosurg Soc 62(4):442–449. https://doi.org/10.3340/jkns.2018.0178
10. Yasaka K, Akai H, Kunimatsu A, Kiryu S, Abe O (2020) Prediction of bone mineral density from computed tomography: application of deep learning with a convolutional neural network. Eur Radiol 30(6):3549–3557. https://doi.org/10.1007/s00330-020-06677-0
11. Hussain D, Han SM (2019) Computer-aided osteoporosis detection from DXA imaging. Comput Methods Prog Biomed 173:87–107. https://doi.org/10.1016/j.cmpb.2019.03.011
12. Xiao P, Zhang T, Dong XN, Han Y, Huang Y, Wang X (2020) Prediction of trabecular bone architectural features by deep learning models using simulated DXA images. Bone Rep 13:100295. https://doi.org/10.1016/j.bonr.2020.100295
13. Mohamed EI, Meshref RA, Abdel-Mageed SM, Moustafa MH, Badawi MI, Darwish SH (2019) A novel morphological analysis of DXA-DICOM images by artificial neural networks for estimating bone mineral density in health and disease. J Clin Densitom 22(3):382–390. https://doi.org/10.1016/j.jocd.2018.08.006
14. Meng J, Sun N, Chen Y, Li Z, Cui X, Fan J, Cao H, Zheng W, Jin Q, Jiang L, Zhu W (2019) Artificial neural network optimizes self-examination of osteoporosis risk in women. J Int Med Res 47(7):3088–3098. https://doi.org/10.1177/0300060519850648
15. Kanis JA, Harvey NC, Johansson H, Odén A, McCloskey EV, Leslie WD (2017) Overview of fracture prediction tools. J Clin Densitom 20(3):444–450. https://doi.org/10.1016/j.jocd.2017.06.013
16. De Vries BCS, Hegeman JH, Nijmeijer W, Geerdink J, Seifert C, Groothuis-Oudshoorn CGM (2020) Comparing three machine learning approaches to design a risk assessment tool for future fractures: predicting a subsequent major osteoporosis fracture in fracture patients with osteopenia and osteoporosis. Osteoporos Int (this issue)
17. Benjamens S, Dhunnoo P, Meskó B (2020) The state of artificial intelligence-based FDA-approved medical devices and algorithms: an online database. NPJ Digital Med 3:118. https://doi.org/10.1038/s41746-020-00324-0
18. Nagendran M, Chen Y, Lovejoy CA, Gordon AC, Komorowski M, Harvey H, Topol EJ, Ioannidis JPA, Collins GS, Maruthappu M (2020) Artificial intelligence versus clinicians: systematic review of design, reporting standards, and claims of deep learning studies. BMJ (Clin Res Ed) 368:m689. https://doi.org/10.1136/bmj.m689

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.