EOR | volume 5 | October 2020
DOI: 10.1302/2058-5241.5.190092
www.efortopenreviews.org

EFORT open reviews

Artificial intelligence in orthopaedics: false hope or not? A narrative review along the line of Gartner's hype cycle

Jacobien H.F. Oosterhoff[1,2]
Job N. Doornberg[1,3]
Machine Learning Consortium[4]

- Artificial Intelligence (AI) in general, and Machine Learning (ML)-based applications in particular, have the potential to change the scope of healthcare, including orthopaedic surgery.
- The greatest benefit of ML is in its ability to learn from real-world clinical use and experience, and thereby its capability to improve its own performance.
- Many successful applications are known in orthopaedics, but have yet to be adopted and evaluated for accuracy and efficacy in patients' care and doctors' workflows.
- The recent hype around AI triggered hope for development of better risk stratification tools to personalize orthopaedics in all subsequent steps of care, from diagnosis to treatment.
- Computer vision applications for fracture recognition show promising results to support decision-making, overcome bias, process high-volume workloads without fatigue, and hold the promise of even outperforming doctors in certain tasks.
- In the near future, AI-derived applications are very likely to assist orthopaedic surgeons rather than replace us. 'If the computer takes over the simple stuff, doctors will have more time again to practice the art of medicine'.[76]

Keywords: artificial intelligence; computer vision; data-driven medicine; machine learning; orthopaedic surgery; orthopaedic trauma; personalized medicine; prediction tools

Cite this article: *EFORT Open Rev* 2020;5:593-603.
DOI: 10.1302/2058-5241.5.190092

Introduction

Artificial intelligence (AI) is believed to have the capacity to change the scope of medicine, much as the introduction of smartphones changed our day-to-day lives. AI and machine learning (ML) are terms commonly used to cover a range of computer applications such as ML-derived clinical decision support, deep learning (DL)-based computer vision and natural language processing (NLP). In essence, computers use human-created algorithms for analysing patterns in data and improve their performance by learning from their own mistakes. The increase in (cheap) powerful computers and availability of larger and more robust data have fuelled the use of ML in healthcare.[1]

For decades, data-driven algorithms have been showing promising results as valuable diagnostic tools to assist clinicians in many respective specialties. As early as the 1980s a data-driven clinical prediction tool to determine which patients with chest pain presenting to the ED (emergency department) could be safely discharged home versus patients who were at high risk of myocardial infarction requiring admission to the intensive care unit (ICU)[2,3] overcame doctors' inconsistent and inefficient admission strategies. This greatly improved workflow in the ED and resulted in fewer admissions while improving patients' outcomes. Now, 30 years later, many hospitals build on similar clinical prediction tools and conduct data-driven algorithms to improve workflow from simple tasks in EDs to complex decision-making in ICUs.[4] In the era of AI, these data-driven algorithms are augmented with ML with two theoretical benefits: (1) to add non-linear correlations to the models; and (2) eventually to become self-learning to improve performance.

However, according to the Gartner hype cycle,[5] we are over the top of the curve and coming down the slope to realize that AI is not going to solve all patients' and doctors' problems (Fig. 1). Nevertheless, many successful applications are known: computer vision DL models screen over 50,000 mammograms annually for breast cancer in the Massachusetts General Hospital in Boston.[6] In orthopaedics, our Massachusetts General Hospital-based SORG (Skeletal Oncology Research Group) is on the

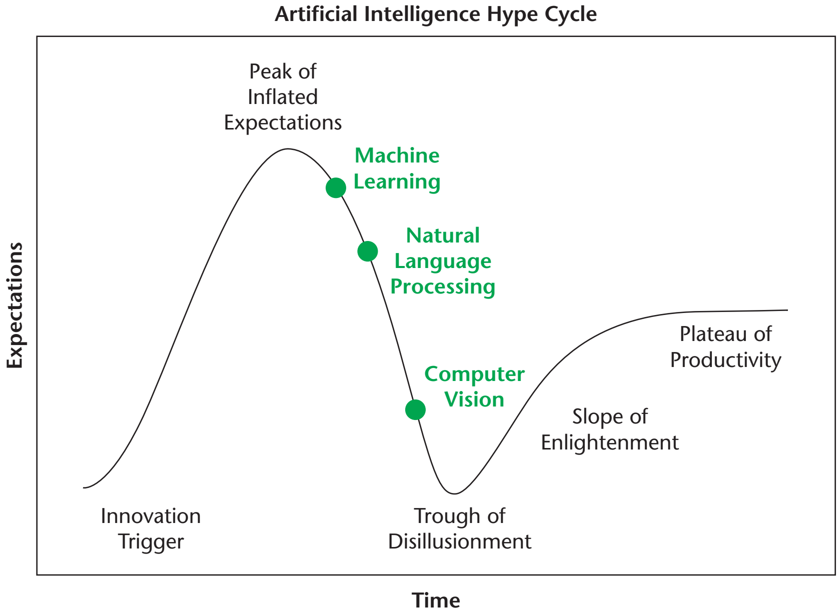

Fig. 1 Artificial Intelligence Hype Cycle, Machine learning, Natural Language Processing and Computer Vision on its way down – Adapted from Gartner Hype Cycle for Artificial Intelligence, 2019 gartner.com/smarterwithgartner.

frontier of ML in orthopaedic musculoskeletal oncology to provide advanced models for predicting surgical outcomes to improve patient-centred care,[7] and the Traumaplatform ML Consortium is broadening the scope of AI to orthopaedic trauma.[8–10] However, critics may argue: 'Why do so many promising applications have yet to be adopted in patients' clinical care or doctors' workflow?'

In this narrative review we focus on AI in orthopaedic surgery. We use respective examples of factual ML applications in orthopaedics to illustrate the great potential of AI to assist orthopaedic surgeons. In contrast, we will present methodologically sound ML applications, which have not, to date, made it to clinical practice, to exemplify AI's shortcomings. Finally, we will present a practical stepwise approach on how to develop, validate, test and implement a ML application by using an example of a clinical prediction tool for discharge destination of patients with hip fractures. All examples serve as a narrative along the line of Gartner's hype cycle to explore the question: 'Artificial intelligence in orthopaedics: false hope or not?'

Part I: orthopaedic surgery is all about risk stratification – how can AI assist?

Risk of bias: risk stratification to neutralize the influence of 'biased' surgeons

In orthopaedics – although generalizable risk factors are well known – the probability of a favourable outcome *or* an adverse event for each respective individual patient that we care for (i.e. risk stratification), is currently still at best an educated guess when taking into account the great number of all unique specific patient and injury characteristics. ML-derived prediction models – that ultimately become self-learning and thereby constantly increase in accuracy – have great clinical potential in such risk stratification. This is based on the premise that high quality data are (prospectively) collected for the specific prediction-task at hand.[11] For example, predicting which elderly hip fracture patient has high probability of developing delirium on admission means they can be targeted for preventive measures.[12] Another example of a 'non-medical' risk stratification, but merely a useful logistical risk stratification that improves 'workflow' in the same frail patient group, is predicting discharge destination which could reduce expensive hospital admission days by streamlining post-operative pathways.[13]

In the era of data-driven care and personalized or 'precision' medicine, decision-making in orthopaedic trauma surgery is flawed by selection bias of surgeons because we still lack good quality prospective outcome data for many common injuries. Moreover, surgeons are notoriously poor at accurately predicting patient outcomes.[14] Hence, there is great – undesired – variation in treatment. For example, in the Netherlands 20% of patients with a wrist fracture undergo surgery, while 80% of patients in Australia are treated operatively.[15,16] Consensus on the optimal treatment strategies for such common fractures is lacking and this leads to sub-optimal workflow, physical impairment and unnecessary costs. When aiming for global consensus, global collective – open access – use of available data is needed.

First, combining data from multiple institutions is difficult because of ethical, legal, political and administrative barriers. However, it can be done: intensive care doctors showcased an innovative example of the Medical Information Mart for Intensive Care (MIMIC) database that was developed by the Massachusetts Institute of Technology (MIT) and Beth Israel Deaconess Medical Centre (BI), Boston MA, USA.[17] It is free to use and to develop prediction tools for non-commercial purposes. This 'open access' mentality will allow improved patient care in a collaborative effort, rather than multiple fragmented individual efforts around the world that have very poor external validity. As such, slowly, supporting systems for collective use of data have come across into healthcare, but have not been evaluated in orthopaedics other than in our implant registries. Distributed learning training – an algorithm learning from data without data leaving the hospital[18] – has been proven an effective alternative for sharing patient imaging data in other specialties for computer vision.[19] Data are allowed to be kept at the source where they are easier to handle and secure.

Some firms – start-ups and for-profit organizations – are bypassing hospital routes to buy data directly from patients in order to receive identified data for model development. In the upward slope of Gartner's hype cycle, over 90 well-funded AI-driven imaging and diagnostic solutions start-ups have been founded to date with combined funding of $1.5 billion.[20] There is a great challenge in conducting infrastructure and pathways for efficient high-quality data-sharing and feasible model development in orthopaedics globally. The benefits of sharing data have been recognized by governments and intergovernmental organizations around the world to promote transparency, accountability and value creation by making data available to all.[21] When data are stored centrally on servers, we can aim for 'open access' anonymized safe data-sharing and applications and thus aim for personalizing orthopaedic care globally and accessibly throughout the world.[17]

Second, we should be cautious when combining data. Combined data can be used when data were collected for a specific research question and collected in an appropriate representative way. In particular, differences in healthcare systems should be acknowledged when combining and translating data through various countries. For example, our discharge prediction tool for elderly patients with a hip fracture that was deployed in Boston MA on data collected through the United States,[12] will likely not be externally valid in different healthcare systems in the Netherlands or Australia. More research is needed to explore these limitations of AI, in particular for ML-driven prediction tools, or computer vision applications using imaging from dissimilar machines from different parts of the world.

In conclusion, treatment is not only influenced by biased surgeons, but decision-making can be biased by differences in healthcare plans and insurance systems.[22] In the clinical case of predicting discharge position after hip fracture surgery, facilities in the United States are limited by insurance approval, whereas in the Netherlands they are limited by availability. This makes generalizing predicted probabilities difficult; an algorithm should be externally validated thoroughly to overcome these discrepancies, as we will elaborate on below.

Risk of bias due to (lack of) experience: risk stratification based on 'big data'

Junior doctors gain experience by treating hundreds of patients during their training. Senior doctors may be considered experienced after treating thousands of patients. Both are prone to bias:[23] the first due to lack of experience, the latter due to personal subjectivity of one's experience.[24] Based on 'objective experience' with greater than 10,000s of patients, DL-driven computer vision and ML-derived prediction tools could alert clinicians about decisions that are at risk of bias. For example, in terms of decision-making in EDs, the majority of patients are seen by junior doctors. Junior doctors are known to misdiagnose significant trauma abnormalities on radiographs.[25] Food and Drug Administration (FDA)-approved and commercially available computer vision applications[26] can produce a heatmap on a radiograph where there is high probability of suspected fracture to alert the junior doctors and improve risk management. In addition, situations with high cognitive load for clinicians, such as decision-making at the end of a clinic day, could be supported by ML predictions. If non-biased ML predictions and real-life clinician decision-making differ in these situations, clinicians can be alerted.[27]

The common claim for ML-derived prediction tools is that a better decision can be made with a model, than without.[28] Transparency and traceability of the decision-making process of AI systems must be made available to physicians in order to avoid fear of the 'black box': 'How did the computer come to this decision?'. Therefore, it is important for orthopaedic trauma surgeons to have a foundation of knowledge of ML, as well as how it may affect and impact models, in order to critically assess predictions generated by ML and interpret the advice on probability of outcome in clinically meaningful ways. Not only treating physicians, but also patients are becoming important consumers of predictive analytics since patients are included in decisions about their treatments. Therefore, better tools to gain insight into risk stratification and communication to patients are needed to achieve true shared decision-making.[1]

When intended to diagnose, treat or prevent disease, ML-derived applications are defined as a medical device under the Food, Drug, and Cosmetic Act in the United States.[29] In Europe, ML-derived applications are required to be approved by the Medical Device Regulations as defined by the CE Mark.[30] In addition, regulatory US and European platforms are not yet equipped to oversee AI's insertion into medical practice.[29]

Part II: three forms of machine learning to aid clinical decision-making

Natural language processing (NLP)

Natural language processing (NLP) is a field of deep learning (DL), with the ability for a computer to understand and analyse human language. Google translate is the best-known non-medical example. DL is a class of ML characterized by the use of neural networks, in which the algorithm learns to distinguish patterns directly from data and learns on its own to select features to classify the input data. The goal of NLP is to translate the natural human language of a patient's medical record, for example surgery reports, into structured format data to query for the presence or absence of a finding.[31] In orthopaedics, NLP has been applied to identify surgical site infections in free-text notes of medical records and achieved predictive abilities comparable with the manual abstraction process and superior to models that used administrative data only.[32] In hip arthroplasty, NLP has been used to identify common data elements[33] and classification of periprosthetic femur fractures.[34] Our group applied NLP to evaluate unstructured free-text patient-experience reviews of orthopaedic surgeons throughout the United States. Patient experience reflects quality of care from the patient's perspective, hence these are important data that can teach us about what creates an (un)satisfying experience.[35]

Another simple, yet very elegant, application of NLP in clinical practice has been developed at the Beth Israel Deaconess Center (BIDC) by Steven Horng – Emergency Physician and Clinical Lead for ML – and colleagues.[36] In the BIDC's emergency department, the NLP algorithm automatically 'reads' the triage nurse's admission note. Subsequently, it provides a drop-down menu of ICD diagnoses in order of differential diagnostic likelihood – rather than alphabetical – based on written clinical triage data. Moreover, this algorithm is subsequently self-learning based on the final entered ICD diagnosis, increasing the accuracy of the drop-down differential diagnosis based on plain written text.

When debating, 'AI, false hope or not?', one could consider the larger sum of these respective small advances in our clinical workflow to result in a major reduction of time we spend on our computers (Fig. 2).

Clinical prediction rule

Predictive tools in orthopaedics consist of diagnostic as well prognostic outcome applications. In orthopaedic trauma, ML-derived clinical prediction rules may enhance workflow in the ED:[37,38] patients clinically suspected for scaphoid fracture are referred for radiographic evaluation. Of these, up to 20% of patients with a negative radiograph have sustained an actual scaphoid fracture.[39] The developed Clinical Prediction Rule can aid clinicians

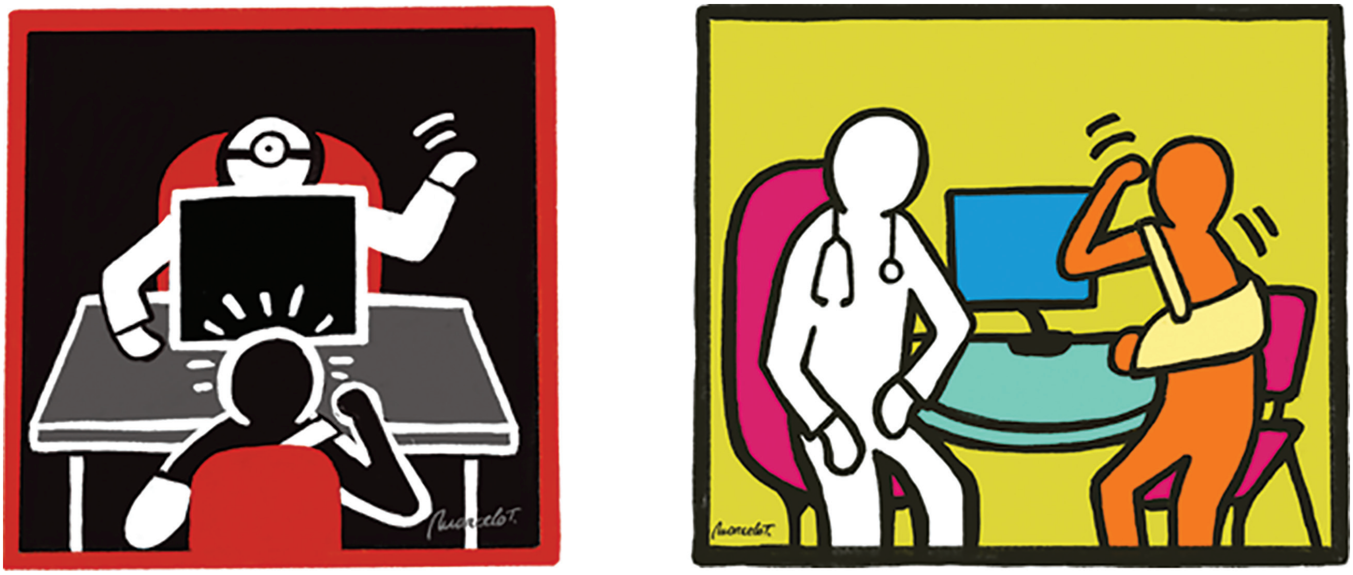

Fig. 2 AI is very likely to assist orthopaedic surgeons: 'If the computer takes over the simple stuff, doctors will have more time again to practice the art of medicine' (Courtesy: Marcello Lavallen).

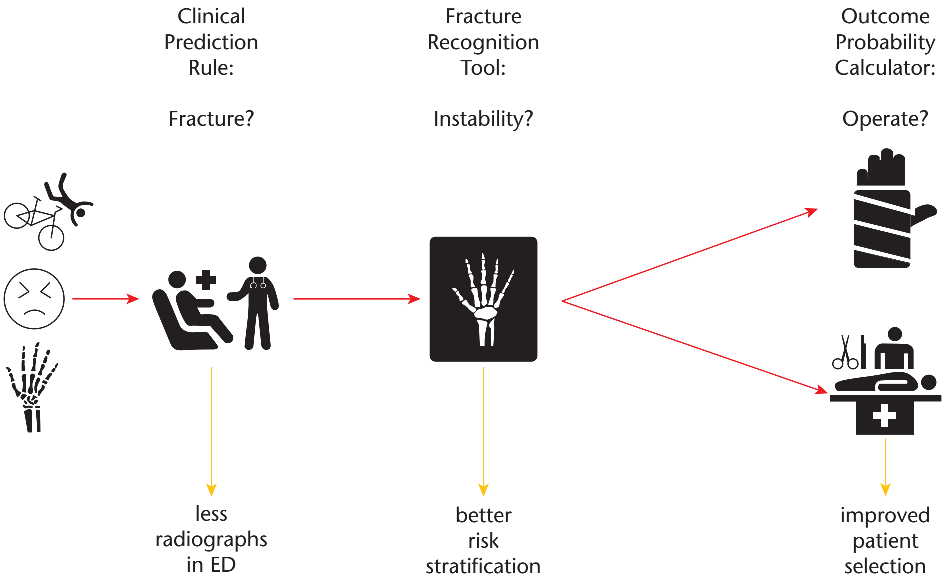
Clinical Prediction Rule:
Fracture?
Fracture Recognition Tool:
Instability?
Outcome Probability Calculator:
Operate?
less radiographs in ED
better risk stratification
improved patient selection

Fig. 3 Workflow for patients clinically suspected for a distal radius fracture.
Note. ED, emergency department.

in identifying patients requiring advanced imaging (i.e. magnetic resonance imaging (MRI) or computed tomography (CT)) and thereby may reduce the number of requested advanced imaging and potential unnecessary casting procedures for up to 31% of patients.[38] Similarly, using the Ottowa Ankle Rules, a combination of predictive clinical parameters increasing the likelihood of a fracture with an additional benefit of its self-learning and correcting capacity, could support and guide the clinician when taking a history and performing a physical examination. Hence, there would be improved risk stratification for advanced imaging of patients with ongoing improved accuracy when results are fed back into the ML algorithm.

Computer vision for fracture recognition

Computer vision is a domain of DL and describes the process of a machine understanding images or videos, and could be useful to aid diagnostic decision-making in fracture care. In computer vision, convolutional neural networks (CNNs) have proven to be effective for these purposes.[40] Using pre-trained CNNs enables us to transfer knowledge to a specific new fracture recognition task, without the need for new time-consuming computational training. Our systematic review addressed the promise and potential utility in fracture care, and found computer vision was nearly as good as and even outperformed humans in detecting certain common fractures.[9] When classifying proximal humerus fractures, often misdiagnosed due to variable presentation, a CNN outperformed general physicians and general orthopaedic surgeons, but with the same performance as specialized upper extremity surgeons. The CNN was trained on ~2000 radiographs classified according to the Neer classification.[41] Moreover, few studies have been published showing that AI performs at a human level in recognizing fractures on plain radiographs taken in the ED of patients with wrist, hand, and ankle injuries with at least 83% accuracy.[42–45] Arguably, these studies all included simple – easy to identify – fractures only.

Of critical note, however, subtle and invisible (occult) fractures may be more challenging than fractures that are easy to detect. In the case of the aforementioned clinically suspected scaphoid fracture, a scaphoid fracture is relatively subtle on radiographs and is often overlooked by non-specialists.[46] Even specialists cannot detect some scaphoid fractures on radiographs – so-called radiographically occult fractures. When applying computer vision to identify true fractures among suspected fractures, many of which were radiographically invisible to human observers, computer vision did not outperform humans. Along Gartner's line: CCN for fracture recognition was embraced for its high potential and lured in many investors supporting numerous start-ups for billions of dollars. But as we are now over the top of the hype cycle, we recognize that, for example, occult fractures of the scaphoid, remain occult for expert surgeons as well as for a specially trained CNN for scaphoid fractures.[47]

This uncovers one of the problems of supervised learning of CNN for musculoskeletal computer vision of occult fractures: training of the algorithm requires a great number of cases, with a reference standard (MRI, CT or follow-up radiographs) which is at best debatable in accuracy. At the stage we are at now, computer vision will miss (occult) scaphoid fractures just as often as orthopaedic surgeons and radiologists do. However, in other specialties, computer vision has been proven to outperform specialists in cancer screening in picking up early tumours that are often missed, even by specialists.[48,49] The hope of computer vision in orthopaedics is early accurate diagnosis and classification, to improve treatment outcomes. At this point, orthopaedic surgeons are on par with AI, as even the first FDA-approved computer vision application in orthopaedics (OsteoDetect) does not exceed specialists' accuracy in detecting and diagnosing distal radius fractures.[26]

Outcome calculator

Risk stratification in orthopaedics has the potential to neutralize the influence of biased surgeons and thus overcome treatment inconsistencies,[16,50] thereby improving patients' functional outcomes and reducing associated healthcare costs (Fig. 3). Thus, small significant changes in daily decision-making in high-volume patient care will result in important overall public health advances.[51,52] In orthopaedics, ML-derived decision tools to assist clinicians in treatment outcomes have been developed in arthroplasty,[53] trauma,[10,12,38] oncology and spinal disorders.[54–57] In orthopaedic oncology, decision tools show accurate performance characteristics in pre-operative estimation of survival in patients with spinal or extremity metastatic disease.[54,55] The developed tools may enhance personalized survival prediction, from 30 days up to five years, and aid shared treatment decision-making, both surgical and non-surgical. In arthroplasty, estimation of patients who will benefit from elective surgery will support optimization in treatment strategy, and prevent patients undergoing an elective procedure with an unacceptably high (individual) risk of adverse events.[55] In orthopaedic trauma, an outcome calculator was developed to identify pre-operative risk of post-operative delirium in hip fracture surgery[12] and the ML algorithm will likely improve the efficiency of a screening programme aimed at identifying patients at risk for delirium. However, the clinical efficacy of the latter tool has yet to be determined and will be the subject of clinical testing and implementation studies.

Although there are many studies on development of decision tools, few authors have driven further development by successful external validation.[58–60] Methods for evaluation and monitoring models to ensure continued accuracy and performance are in their infancy with regard to their imbedding ML in healthcare.[61]

In the final part of this narrative review, we will demonstrate a logical stepwise approach from clinical problem to implementation, derived from a successfully implemented ML application, which is suggested to be followed to ensure quality in orthopaedic ML research.

Part III: stepwise approach from clinical decision-making problem to implementation

The methodology follows the framework for prediction models proposed by Professor Steyerberg et al,[28] and covers the range of development of applications such as NLP, computer vision and clinical decision support as discussed above (Fig. 4).

Step 1. Predictive modelling: development of a machine learning algorithm

Data derived from various study designs addressing the clinical decision-making problem at hand could be used for predictive modelling with the use of ML; retrospective, prospective, registry data and nested case-control studies fit best for prognostic modelling whereas cross-sectional and case-control study design fit better for diagnostic modelling.[62] The benefit of ML may be best realized with larger data sets, particularly those that are periodically updated, with the rule of thumb of having ≥ 200 events and ≥ 200 non-events.[63] For example, a ML algorithm for delirium prediction following elderly hip fracture surgery, and various other SORG ML algorithms, were developed with a large clinical database from the American College of Surgeons (ACS) National Surgical Quality Improvement Programme (NSQIP).[12,56,57]

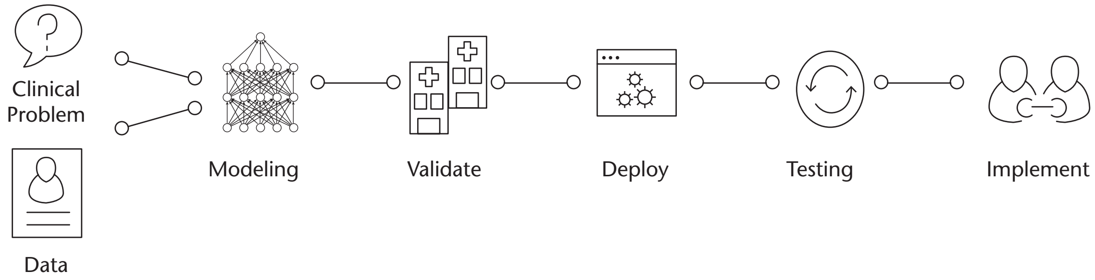

Fig. 4 Flowsheet from clinical problem to implementation.

A function is generated consisting of an outcome variable (dependent variable) which is predicted from a given set of features (independent variables). In the case of development of a clinical prediction rule or outcome calculator, variable importance is first carried out to identify and select those features that contribute most to the outcome variable with clinical importance in mind. Variables included may contain clinical and radiological findings (e.g. patient demographics, trauma mechanism or classification of fracture), as well as intra-operative findings and surgical techniques (e.g. screw and/or plate fixation or arthroplasty). In the case of computer vision and NLP, the algorithm distinguishes patterns directly from data and learns on its own to select features to classify the input data (essentially black boxes – e.g. edges, curves, colour). Training and internal validation of the supervised ML algorithm continues ('run') until the model achieves the best model performance. The delirium hip fracture prediction tool targeted post-operative delirium as the dependent variable, with easy, readily available independent variables derived from variable importance (i.e. age, BMI, ASA class, functional status, pre-operative dementia, pre-operative delirium, pre-operative need for mobility-aid and pre-operative creatinine level).

Predictive performance of ML algorithms is assessed according to Steyerberg's structured stepwise ABCD-approach: calibration-in-the-large, or the model intercept (A); calibration slope (B); discrimination, with a concordance statistic (C); and clinical usefulness, with decision-curve analysis (D).[28,64] In addition, overall model performance – a composite of discrimination – is assessed using the Brier score, compared with the null model Brier score.[65]

Classification algorithms include linear classifiers (logistic regression, naïve Bayes), support vector machine, classification trees or neural networks (Fig. 5). Linear classifiers are easy to interpret, and fast to train. Non-linear classifiers are more flexible and have the ability to capture more complex patterns, but are, in small samples, prone to overfitting. Logistic regression involves fitting an S-shaped probability curve to numerical data by taking the log odds to make predictions about binary events. Classification trees (e.g. gradient boosted machine, random forest) use flowchart-like structures to make decisions, which can be readily understood and visualized. Artificial neural networks are inspired by biological neural networks which mainly use so-called feed-forward neural networks with hidden layers and neurons and which are, in general, data-hungry. Support vector machines are based on the idea of finding a hyperplane in a 3D (kernel) scatterplot that can divide a dataset into two classes, and works in general quite well on smaller datasets. A naïve Bayes machine is a product of probabilities, best visualized as a Venn diagram that shows possible logical relations, works well with smaller datasets, and prefers categorial features over continuous features (where a normal distribution is assumed).

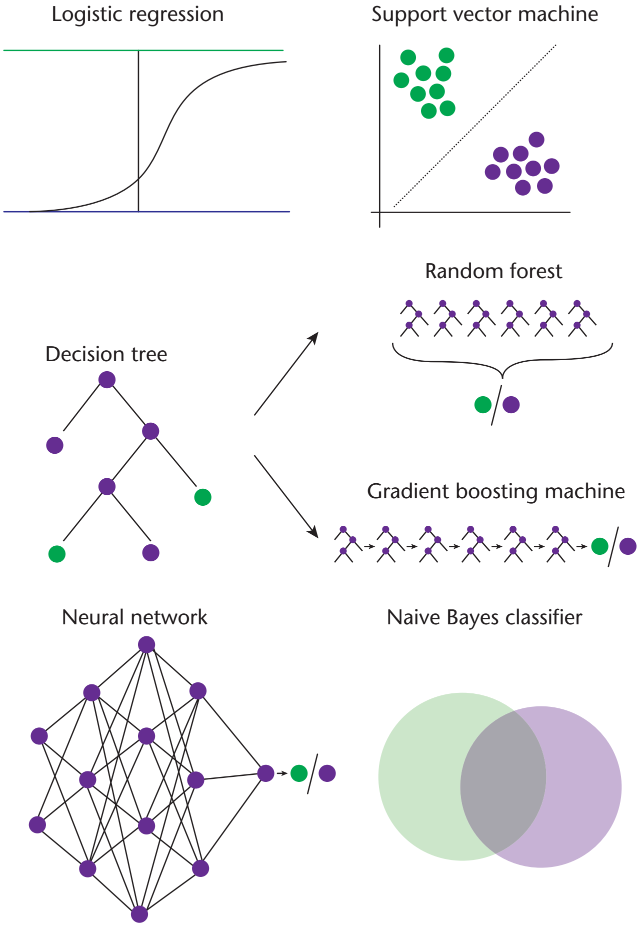

Fig. 5 Classification algorithms. (Courtesy: B.Y. Gravesteijn)

On one hand, there is no one solution when choosing the right ML algorithm. The decision is taken after conducting a research question, preparing data and building various models. Comparing model performance is based on all metrics according to the ABCD approach,[28] combined with the most clinically meaningful variable importance. The ML algorithm development for delirium prediction in hip fracture surgery led to almost perfect model performance combined with clinically meaningful feature importance, outperformed the default strategy of screening all patients, and included easy and readily available variables.

External validation

External validation is essential to assess performance and generalizability of the algorithm before implementation in clinical practice. External validation can be carried out with temporal, geographical or fully independent validation.[65] External validation is also important because model performance might differ across populations, making an unvalidated algorithm less reliable.[60] Of the current few externally validated ML algorithms the validation cohort was derived from retrospective analysis at a large, tertiary care centre.[58–60] A developed, internally validated ML algorithm is applied to a separate validation dataset to assess model performance according to the same metrics as above. Even though various institutions may all be using the same electronic health record (EHR) vendor, the data structure, field meanings and extent of data cleaning likely differ across organizations.[66] For future research, when prospectively collecting data, common data elements for common data models could support combining data from various institutes and validation of prediction models globally[67] and thereby support fully independent validation.[68]

As discussed above, our discharge prediction tool for elderly patients with a hip fracture was deployed in Boston MA on United States data.[12] Differences in healthcare systems, standards of care and treatment strategies can prohibit generalizability to other countries. In some situations, when external validation reveals low generalizability, re-calibration strategies are allowed.[69] In re-calibration, particular components of the developed model are modified and tailored for each study population (such as the intercept of the model or variable effects).[70]

Evaluation and implementation

If found to be externally valid, clinicians might use an available (web) application to help incorporate the algorithm into practice to aid decision-making and target actions to be a priority (e.g. https://sorg-apps.shinyapps.io/hipfxdelirium/). A real-time clinical prediction rule, computer vision model or outcome calculator based on the developed ML algorithm and routinely collected clinical data is best established and validated in EHR systems.[71] Derived predictions are integrated and calculated automatically and made available to the clinician.[71] Although ML is a new methodology that greatly expands the ability to analyse data, implementation should follow the same rules as the previously developed diagnostic test.[72,73]

Efficacy of the developed ML algorithm is ideally assessed through large randomized controlled trials (RCTs).[74] ML-derived decision support has great power to assist clinicians and change the scope of medicine; however, many powerful algorithms are not utilized yet.[66] Consider the following scenario: a patient is scheduled for hip fracture surgery and randomized to either the intervention or control arm of an RCT. In the intervention arm, an intervention is based on high probability derived from the developed ML algorithm. In the control arm, treatment is according to common practice. The proposed primary end-point is incidence of post-operative delirium to determine benefit from the developed ML algorithm and clinical importance (i.e. patient outcomes).

ML requires the use of a computer and EHR integration, which has implications for patient privacy and creates obstacles for implementation.[73] Physicians will need to open the application, enter information and then return to using it in the EHR.[66] The biggest challenge is incorporating an ML-derived decision support tool into an EHR workflow. In addition, the distinctive characteristics of ML-based software require a regulatory approach, allowing necessary steps to improve treatment while ensuring that the algorithm is safe.[75]

Improvement of the algorithm: continuous self-learning

The increase in data set size substantially improves ML model performance, as a response to changes in practice or patient population. Ongoing data collection will lead to improved ML models, though with gradually diminishing returns.[73] The great advantage of ML algorithms over decision rules is the ability to improve accuracy of the model over time, including earlier disease detection, more accurate diagnosis, identification of new observations or patterns, and development of personalized diagnostics and treatment.

Conclusions

Many argue that AI will change the scope of medicine. Indeed, along the upslope of Gartner's hype cycle, $1.5 billion has been invested in AI in healthcare,[20] and counting. However, coming over the top of the hype curve, we recognize the methodological limitations of ML and DL: for example, a computer can recognize an obvious fracture,[41] which may be beneficial as a support tool for junior doctors in an ED under a high demanding workload.[25] But for an occult scaphoid fracture, CNN algorithms have yet to outperform orthopaedic specialists.[47]

On the downward slope of Gartner's line, we come to realize that many promising ML prediction tools and DL image recognition tools have been developed with good intentions for commercial benefit, but very few have been externally validated – systematically tested on accuracy in clinical workflow – or implemented in daily practice to date. To do so in orthopaedics, we face ethical, legal, political and administrative barriers. To move forward along the slope of enlightenment, we strongly argue for collaboration in an 'open access' mentality as intensive care specialists do:[17] share good quality prospective data to improve the accuracy and external validity of AI-derived algorithms; and – in an ideal world – continue prospective data collection with an active feedback loop to improve performance.

We envision the plateau of productivity of the hype cycle as follows: AI-derived applications will facilitate data-driven personalized care for our patients, limiting surgeons' bias, and empower shared decision-making on patient specific data. AI is likely to assist orthopaedic surgeons rather than replace us: 'If the computer takes over the simple stuff, doctors will have more time again to practice the art of medicine'.[76]

AUTHOR INFORMATION

[1]1Department of Orthopaedic Surgery, Amsterdam UMC, University of Amsterdam, the Netherlands.
[2]Department of Orthopaedic Surgery, Massachusetts General Hospital, Harvard Medical School, Boston, MA, USA.
[3]Department of Orthopaedic & Trauma Surgery, Flinders Medical Centre, Flinders University, Adelaide, SA, Australia.
[4]For full details see Appendix 1.

Correspondence should be sent to: Job N. Doornberg, Department of Orthopaedic Trauma Surgery, Flinders Medical Centre (Level 5), GPO Box 2100, Adelaide 5001, SA5042, Australia.
Email: doornberg@traumaplatform.org

ICMJE CONFLICT OF INTEREST STATEMENT

JD reports receipt of a grant from Marti-Keuning Eckhardt Foundation for the submitted work.

The other authors declare no conflict of interest relevant to this work.

FUNDING STATEMENT

No benefits in any form have been received or will be received from a commercial party related directly or indirectly to the subject of this article.

LICENCE

REFERENCES

1. **Peterson ED.** Machine learning, predictive analytics, and clinical practice: can the past inform the present? *JAMA* 2019;Nov 22.

2. **Goldman L, Cook EF, Johnson PA, Brand DA, Rouan GW, Lee TH.** Prediction of the need for intensive care in patients who come to emergency departments with acute chest pain. *N Engl J Med* 1996;334:1498–1504.

3. **Reilly BM, Evans AT, Schaider JJ, et al.** Impact of a clinical decision rule on hospital triage of patients with suspected acute cardiac ischemia in the emergency department. *JAMA* 2002;288:342–350.

4. **McWilliams CJ, Lawson DJ, Santos-Rodriguez R, et al.** Towards a decision support tool for intensive care discharge: machine learning algorithm development using electronic healthcare data from MIMIC-III and Bristol, UK. *BMJ Open* 2019;9:e025925.

5. **Car J, Sheikh A, Wicks P, Williams MS.** Beyond the hype of big data and artificial intelligence: building foundations for knowledge and wisdom. *BMC Med* 2019;17:143.

6. **Lehman CD, Yala A, Schuster T, et al.** Mammographic breast density assessment using deep learning: clinical implementation. *Radiology* 2019;290:52–58.

7. **Skeletal Oncology Research Group.** Machine learning for the practicing surgeon. https://www.sorg-ai.com/ (date last accessed 1 November 2019).

8. **Traumaplatform AI.** Orthopaedic trauma: artifical intelligence. https://www.traumaplatformai.org/ (date last accessed 10 December 2019).

9. **Langerhuizen DWG, Janssen SJ, Mallee WH, et al.** What are the applications and limitations of artificial intelligence for fracture detection and classification in orthopaedic trauma imaging? A systematic review. *Clin Orthop Relat Res* 2019;477:2482–2491.

10. **Hendrickx LAM, Sobol GL, Langerhuizen D, et al.** A machine learning algorithm to predict the probability of (occult) posterior malleolar fractures associated with tibial shaft fractures to guide 'malleolus first' fixation. *J Orthop Trauma* 2019; doi: 10.1097/BOT.0000000000001663 [Epub ahead of print].

11. **Ewald H, Ioannidis JPA, Ladanie A, Mc Cord K, Bucher HC, Hemkens LG.** Nonrandomized studies using causal-modeling may give different answers than RCTs: a meta-epidemiological study. *J Clin Epidemiol* 2019;118:29–41.

12. **Oosterhoff JHF, Karhade AV, Oberai T, Doornberg JN, Schwab JH.** Development of machine learning algorithms for prediction of postoperative delirium in elderly hip fracture patients. Manuscript submitted for publication to *Clin Orthop Relat Res* 2019.

13. **Pitzul KB, Wodchis WP, Kreder HJ, Carter MW, Jaglal SB.** Discharge destination following hip fracture: comparative effectiveness and cost analyses. *Arch Osteoporos* 2017;12:87.

14. **Bloembergen CHMD, van de Graaf VA, Virgile A, et al.** Infographic: can even experienced orthopaedic surgeons predict who will benefit from surgery when patients present with degenerative meniscal tears? A survey of 194 orthopaedic surgeons who made 3880 predictions. *Br J Sports Med* 2019;bjsports-2019-101502.

15. **Ansari U, Adie S, Harris IA, Naylor JM.** Practice variation in common fracture presentations: a survey of orthopaedic surgeons. *Injury* 2011;42:403–407.

16. **Walenkamp MMJ, Mulders MAM, Goslings JC, Westert GP, Schep NWL.** Analysis of variation in the surgical treatment of patients with distal radial fractures in the Netherlands. *J Hand Surg Eur Vol* 2017;42:39–44.

17. **Johnson AEW, Pollard TJ, Shen L, et al.** MIMIC-III, a freely accessible critical care database. *Sci Data* 2016;3:160035.

18. **Jochems A, Deist TM, van Soest J, et al.** Distributed learning: developing a predictive model based on data from multiple hospitals without data leaving the hospital: a real life proof of concept. *Radiother Oncol* 2016;121:459–467.

19. **Chang K, Balachandar N, Lam C, et al.** Distributed deep learning networks among institutions for medical imaging. *J Am Med Inform Assoc* 2018;25:945–954.

20. **CB Insights.** From drug R&D to diagnostics: 90+ artificial intelligence startups in healthcare. https://www.cbinsights.com/research/artificial-intelligence-startups-healthcare/ (date last accessed 8 December 2019).

21. **World Health Organization.** WHO data policy. https://www.who.int/publishing/datapolicy/en/ (date last accessed 20 November 2019).

22. **Ridic G, Gleason S, Ridic O.** Comparisons of health care systems in the United States, Germany and Canada. *Mater Sociomed* 2012;24:112–120.

23. **Kahneman D.** *Thinking fast, thinking slow*. New York City: Farrar, Straus and Giroux, 2011.

24. **Harris I.** *Surgery, the ultimate placebo*. Randwick: University of New South Wales Press, 2016.

25. **McLauchlan CA, Jones K, Guly HR.** Interpretation of trauma radiographs by junior doctors in accident and emergency departments: a cause for concern? *J Accid Emerg Med* 1997;14:295–298.

26. Voelker R. Diagnosing fractures with AI. *JAMA* 2018;320:23.

27. Parikh RB, Teeple S, Navathe AS. Addressing bias in artificial intelligence in health care. JAMA 2019;Nov 22.

28. Steyerberg EW, Vergouwe Y. Towards better clinical prediction models: seven steps for development and an ABCD for validation. *Eur Heart J* 2014;35:1925–1931.

29. US Food and Drug Administration. Proposed regulatory framework for modifications to artificial intelligence/machine learning (AI/ML)-based software as a medical device (SaMD). https://www.fda.gov/media/122535/download (date last accessed 29 November 2019).

30. Toh TS, Dondelinger F, Wang D. Looking beyond the hype: applied AI and machine learning in translational medicine. *EBioMedicine* 2019;47:607–615.

31. Cai T, Giannopoulos AA, Yu S, et al. Natural language processing technologies in radiology research and clinical applications. *Radiographics* 2016;36:176–191.

32. Thirukumaran CP, Zaman A, Rubery PT, et al. Natural language processing for the identification of surgical site infections in orthopaedics. *J Bone Joint Surg Am* 2019;101:2167–2174.

33. Wyles CC, Tibbo ME, Fu S, et al. Use of natural language processing algorithms to identify common data elements in operative notes for total hip arthroplasty. *J Bone Joint Surg Am* 2019;101:1931–1938.

34. Tibbo ME, Wyles CC, Fu S, et al. Use of natural language processing tools to identify and classify periprosthetic femur fractures. *J Arthroplasty* 2019;34:2216–2219.

35. Langerhuizen D, Brown L, Doornberg J, et al. Analysis of online rating of orthopaedic surgeons using natural language processing. Manuscript submitted for publication to *Clin Orthop Relat Res* 2019.

36. Rotmensch M, Halpern Y, Tlimat A, Horng S, Sontag D. Learning a health knowledge graph from electronic medical records. *Sci Rep* 2017;7:5994.

37. Mulders MAM, Walenkamp MMJ, Slaar A, et al. Implementation of the Amsterdam Pediatric Wrist Rules. *Pediatr Radiol* 2018;48:1612–1620.

38. Bulstra A, Court-Brown C, Doornberg J, et al. Machine learning algorithm to estimate the probability of a true scaphoid fracture among patients with radial wrist pain and tenderness after a fall. Manuscript submitted for publication to *Clin Orthop Relat Res* 2019.

39. Mallee WH, Wang J, Poolman RW, et al. Computed tomography versus magnetic resonance imaging versus bone scintigraphy for clinically suspected scaphoid fractures in patients with negative plain radiographs. *Cochrane Database Syst Rev* 2015;:CD010023.

40. Chartrand G, Cheng PM, Vorontsov E, et al. Deep learning: a primer for radiologists. *Radiographics* 2017;37:2113–2131.

41. Chung SW, Han SS, Lee JW, et al. Automated detection and classification of the proximal humerus fracture by using deep learning algorithm. *Acta Orthop* 2018;89:468–473.

42. Lindsey R, Daluiski A, Chopra S, et al. Deep neural network improves fracture detection by clinicians. *Proc Natl Acad Sci U S A* 2018;115:11591–11596.

43. Amtmann D, Kim J, Chung H, et al. Comparing CESD-10, PHQ-9, and PROMIS depression instruments in individuals with multiple sclerosis. *Rehabil Psychol* 2014;59: 220–229.

44. Olczak J, Fahlberg N, Maki A, et al. Artificial intelligence for analyzing orthopedic trauma radiographs. *Acta Orthop* 2017;88:581–586.

45. Kim DH, MacKinnon T. Artificial intelligence in fracture detection: transfer learning from deep convolutional neural networks. *Clin Radiol* 2018;73:439–445.

46. Suh N, Grewal R. Controversies and best practices for acute scaphoid fracture management. *J Hand Surg Eur Vol* 2018;43:4–12.

47. Langerhuizen DWG, Bulstra AEJ, Janssen SJ et al. Is deep learning on par with human observers for detection of radiographically visible and occult fractures of the scaphoid? In press, *Clin Orthop Relat Res* 2020.

48. Ardila D, Kiraly AP, Bharadwaj S, et al. End-to-end lung cancer screening with three-dimensional deep learning on low-dose chest computed tomography. *Nat Med* 2019;25:954–961.

49. Rodriguez-Ruiz A, Lång K, Gubern-Merida A, et al. Stand-alone artificial intelligence for breast cancer detection in mammography: comparison with 101 radiologists. *J Natl Cancer Inst* 2019;111:916–922.

50. Walenkamp MMJ, Bentohami A, Slaar A, et al. The Amsterdam Wrist Rules: the multicenter prospective derivation and external validation of a clinical decision rule for the use of radiography in acute wrist trauma. *BMC Musculoskelet Disord* 2015;16:389.

51. Mulders MAM, van Eerten PV, Goslings JC, Schep NWL. Non-operative treatment of displaced distal radius fractures leads to acceptable functional outcomes, however at the expense of 40% subsequent surgeries. *Orthop Traumatol Surg Res* 2017;103:905–909.

52. Mulders MAM, Walenkamp MMJ, van Dieren S, Goslings JC, Schep NWL; VIPER Trial Collaborators. Volar plate fixation versus plaster immobilization in acceptably reduced extra-articular distal radial fractures: a multicenter randomized controlled trial. *J Bone Joint Surg Am* 2019;101:787–796.

53. Fontana MA, Lyman S, Sarker GK, Padgett DE, MacLean CH. Can machine learning algorithms predict which patients will achieve minimally clinically important differences from total joint arthroplasty? *Clin Orthop Relat Res* 2019;477:1267–1279.

54. Karhade AV, Thio QCBS, Ogink PT, et al. Predicting 90-day and 1-year mortality in spinal metastatic disease: development and internal validation. *Neurosurgery* 2019;85:E671–E681.

55. Thio QCBS, Karhade AV, Bindels B, et al. Development and internal validation of machine learning algorithms for preoperative survival prediction of extremity metastatic disease. *Clin Orthop Relat Res* 2020 Feb;478(2):322-333.

56. Ogink PT, Karhade AV, Thio QCBS, et al. Development of a machine learning algorithm predicting discharge placement after surgery for spondylolisthesis. *Eur Spine J* 2019;28:1775–1782.

57. Karhade AV, Thio QCBS, Ogink PT, et al. Development of machine learning algorithms for prediction of 30-day mortality after surgery for spinal metastasis. *Neurosurgery* 2019;85:E83–E91.

58. Karhade AV, Ahmed AK, Pennington Z, et al. External validation of the SORG 90-day and 1-year machine learning algorithms for survival in spinal metastatic disease. *Spine J* 2020;20:14–21.

59. Stopa BM, Robertson FC, Karhade AV, et al. Predicting nonroutine discharge after elective spine surgery: external validation of machine learning algorithms. *J Neurosurg Spine* 2019;Jul:1–6.

60. Bongers MER, Thio QCBS, Karhade AV, et al. Does the SORG algorithm predict 5-year survival in patients with chondrosarcoma? An external validation. *Clin Orthop Relat Res* 2019;477:2296–2303.

61. Sendak M, Gao M, Nichols M, Lin A, Balu S. Machine learning in health care: a critical appraisal of challenges and opportunities. *EGEMS (Wash DC)* 2019;7:1.

62. **Steyerberg EW.** Study design for prediction modeling. In: *Clinical prediction models: a practical approach to development, validation, and updating*. New York City: Springer, 2019:37–56.

63. **Van Calster B, Nieboer D, Vergouwe Y, De Cock B, Pencina MJ, Steyerberg EW.** A calibration hierarchy for risk models was defined: from utopia to empirical data. *J Clin Epidemiol* 2016;74:167–176.

64. **Van Calster B, Vickers AJ.** Calibration of risk prediction models: impact on decision-analytic performance. *Med Decis Making* 2015;35:162–169.

65. **Steyerberg EW.** Validation of prediction models. In: *Clinical prediction models: a practical approach to development, validation, and updating*. New York City: Springer, 2019:309–323.

66. **Haas D, Makhni EC, Schwab JH, Halamka JD.** 3 myths about machine learning in health care. *Harvard Business Review*. https://hbr.org/2019/11/3-myths-about-machine-learning-in-health-care (date last accessed 5 December 2019).

67. **Prosperi M, Min JS, Bian J, Modave F.** Big data hurdles in precision medicine and precision public health. *BMC Med Inform Decis Mak* 2018;18:139.

68. **Brown JS, Holmes JH, Shah K, Hall K, Lazarus R, Platt R.** Distributed health data networks: a practical and preferred approach to multi-institutional evaluations of comparative effectiveness, safety, and quality of care. *Med Care* 2010;48: S45–S51.

69. **Debray TPA, Vergouwe Y, Koffijberg H, Nieboer D, Steyerberg EW, Moons KGM.** A new framework to enhance the interpretation of external validation studies of clinical prediction models. *J Clin Epidemiol* 2015;68:279–289.

70. **Riley RD, Ensor J, Snell KIE, et al.** External validation of clinical prediction models using big datasets from e-health records or IPD meta-analysis: opportunities and challenges. *BMJ* 2016;353:i3140.

71. **Meyer A, Zverinski D, Pfahringer B, et al.** Machine learning for real-time prediction of complications in critical care: a retrospective study. *Lancet Respir Med* 2018;6:905–914.

72. **Guyatt GH, Haynes RB, Jaeschke RZ, et al; Evidence-Based Medicine Working Group.** Users' guides to the medical literature: XXV. Evidence-based medicine: principles for applying the users' guides to patient care. *JAMA* 2000;284:1290–1296.

73. **Liu Y, Chen PC, Krause J, Peng L.** How to read articles that use machine learning: users' guides to the medical literature. *JAMA* 2019;322:1806–1816.

74. **Auleley G-R, Ravaud P, Giraudeau B, et al.** Implementation of the Ottawa ankle rules in France: a multicenter randomized controlled trial. *JAMA* 1997;277:1935–1939.

75. **Hwang TJ, Kesselheim AS, Vokinger KN.** Lifecycle regulation of artificial intelligence- and machine learning-based software devices in medicine. *JAMA* 2019. doi:10.1001/jama.2019.16842 [Epub ahead of print].

76. **Ring D.** Can health technology increase compassion? 2018. https://panelpicker.sxsw.com/vote/85251 (date last accessed 30 November 2019).

Appendix 1

Members of the Machine Learning Consortium:
Paul Algra; Michel van den Bekerom; Mohit Bhandari; Michiel Bongers; Charles Court-Brown; Anne-Eva Bulstra; Geert Buijze; Sofia Bzovsky; Neil Chen; Job Doornberg; Andrew Duckworth; J. Carel Goslings; Benjamin Gravesteijn; Olivier Groot; Gordon Guyatt; Laurent Hendrickx; Dirk-Jan Hofstee; Frank IJpma; Ruurd Jaarsma; Stein Janssen; Paul Jutte; Aditya Karhade; Lucien Keijser; Gino Kerkhoffs; David Langerhuizen; Jonathan Lans; Wouter Mallee; Matthew Moran; Margaret McQueen; Marjolein Mulders; Miryam Obdeijn; Tarandeep Oberai; Jacobien H.F. Oosterhoff; Rudolf Poolman; David Ring; Paul Tornetta III; Joseph Schwab; Emil H. Schemitsch; Niels Schep; Inger Schipper; Bram Schoolmeesters; Marc Swiontkowski; David Sanders; Sheila Sprague; Ewout Steyerberg; Stephen D. Walter; Monique Walenkamp.