AI in the treatment of fertility: key considerations

Jason Swain¹ · Matthew Tex VerMilyea² · Marcos Meseguer³ · Diego Ezcurra⁴ · Fertility AI Forum Group

Received: 15 July 2020 / Accepted: 13 September 2020 / Published online: 29 September 2020
© Springer Science+Business Media, LLC, part of Springer Nature 2020

Abstract

Artificial intelligence (AI) has been proposed as a potential tool to help address many of the existing problems related with empirical or subjective assessments of clinical and embryological decision points during the treatment of infertility. AI technologies are reviewed and potential areas of implementation of algorithms are discussed, highlighting the importance of following a proper path for the development and validation of algorithms, including regulatory requirements, and the need for ecosystems containing enough quality data to generate it. As evidenced by the consensus of a group of experts in fertility if properly developed, it is believed that AI algorithms may help practitioners from around the globe to standardize, automate, and improve IVF outcomes for the benefit of patients. Collaboration is required between AI developers and healthcare professionals to make this happen.

Keywords fertility · AI · algorithms · embryos

Introduction

The treatment of infertility has continually evolved since the birth of Louise Brown in 1978. However, a constant factor during this evolution has remained; that success in assisted reproductive technologies (ART) is in the details. Indeed, ART success relies on a complex series of interrelated events, with each procedural step having several variables and impacting the other. Therefore, to excel in ART and achieve the ultimate goal of a healthy singleton live birth in the shortest time span possible, it is necessary to optimize the success and efficiency in each of these procedural steps and variables.

Despite continued advancements in the development of new drugs, innovative stimulation protocols, and lab technologies, success rates continue to still tend in only average around one-third of patients taking home a baby [1].

Notably, reported success rates of some centers are considerably higher than others. These significant differences in ART outcomes observed between countries [2, 3] and even individual clinics [4] are likely a result of patient variability and the minute details of these ART procedural steps.

In recent years, the development and implementation of artificial intelligence (AI) technology have shown the potential to address inefficiencies in various steps of ART, including the standardization of some IVF laboratory processes and particularly in embryo selection. Indeed, AI is also being proposed for clinical applications including diagnostics and precise treatment paths in combination with digital devices remotely collecting real-time data to be analyzed.

What is AI?

Simply put, AI is widely described as a machine behaving in a way that would be called “intelligent” if viewed by a human. It is a broad term that includes machine learning (ML) and deep learning (DL). AI refers to any program with the ability to solve problems, learn from experiences, and perform tasks as human beings normally do. *Machine learning* is a subset of AI that includes statistical techniques which allow machines to improve on experienced tasks, learning from data without a constant supervision [5]. ML algorithms are developed to identify patterns in given data to make predictions and

✉ Diego Ezcurra
diego.ezcurra@emdserono.com

¹ CCRM Fertility Network, Lone Tree, CO, USA

² Ovation Fertility and Texas Fertility Centers, Austin and San Antonio, TX, USA

³ Instituto Valenciano de Infertilidad (IVI) Valencia, INCLIVA-Universidad de Valencia, Valencia, Spain

⁴ EMD Serono, One Technology Place, Rockland MA02370, USA

support decisions. Lastly, *deep learning* is a subset of machine learning that uses multilayer neural networks inspired by the human brain to learn from large amounts of data. In ART, clinical data, patient demographics, and laboratory data can be used to train models to support clinicians and embryologists to answer multiple questions like predicting a stimulation protocol [6], the probability of an oocyte to reach blastocyst stage, the quality of the blastocysts obtained in a treatment, and the likelihood to achieve a pregnancy [7] or eventually live birth.

Artificial intelligence approach in medicine and ART

The recent increase in computer power and data availability with the developments in machine learning has led to an augment of predictive algorithms in most medical fields [8]. While the beginnings of AI dates back to the early twentieth century, from the concepts presented by Alan Turing [9, 10] and Warren McCulloch and Walter Pitts [11], it was not until 2012 that the ML was widely accepted as a viable form of AI [12]. Most health professionals (72%) show great interest in implementing AI in their work, claiming that this technology is an extension, not the extinction of professional capacity in health care [13]. In recent years, there has been significant growth in the number of AI publications in medicine and health, with 61.6% of the papers dated between 2008 and 2017 [14], making it one of the most promising fields of study in health science. In just 1 year, publications on this subject have increased sevenfold [5]. The American Society for Reproductive Biology (ASRM) [15] and European Society for Human Reproduction and Embryology (ESHRE) [8] have already reported on different applications of machine learning ranging from sperm identification and morphology, identifying empty follicles, predicting the stages of embryonic cells and the formation of blastocysts from oocytes, assessing the quality of human blastocysts, predicting live births following embryo transfer, improving the selection of embryos, and defining optimal IVF stimulation protocols [5]. Until July 2019, the FDA (Food and Drug Administration) had approved 39 algorithms based on AI in medicine [16]. Most of those algorithms are for pattern recognition and suggest dichotomous (yes/no) diagnostic decisions, leading to sporadic implementation and use to support clinical and therapeutic decisions (some of them are summarized in Table 1). None of these tools are broadly available and only work on selected platforms or setups. It has been shown that predictive models built on data and domain-specific knowledge can support and improve or solve specific problems. Those examples can be used to inspire developments of similar approaches in different fields [16, 44]. It has been shown that deep learning algorithms have performances similar to the healthcare professional, and also it has been reported that ML applications might

perform better than healthcare professionals, but validation and prospective clinical studies are necessary to confirm those results [44]. In conclusion, AI-enabled decision support algorithms allow medical teams to make more accurate diagnoses [13]. However, it is important to consider the limitations of those studies: most of them assess diagnostic accuracy and performances of models in isolation, thus not reflecting clinical practice. A major limitation to the interpretation of claims reporting a superior performance of these algorithms versus humans is that analytics are performed on previously generated data *in silico*, not prospectively in real-world clinical conditions [16].

Furthermore, the term “validation” about the testing of AI algorithms is used in multiple ways. In a survey of 400 AI papers presented at major conferences, just 6% included code for the papers’ algorithms and 30% included test data and 54% included pseudocode which is a limited summary of an algorithm. Reproducibility is especially challenging in ML for health applications because of the availability, quality, and consistency of clinical or biomedical data. Specifically, in our field, infertility is increasingly affecting couples, and the interest in assisted reproduction technologies is growing. There are a multitude of AI solutions for patients that are only related to conception and pregnancy. Even though many researchers and commercial companies are developing potential AI IVF/ART tools for fertility, there are currently very few products or broadly accepted tools under development that have consistently provided clinical improvement, and many of them cannot be utilized without restrictions, limiting clinical implementation (some examples are summarized in Table 2).

Notwithstanding the large number of existing AI publications, there is no standardization of the required material, and there are no clear minimum criteria for the acceptance of AI and ML studies. This creates a conundrum in terms of the literature and trying to ensure high quality data. Indeed, due to this issue, in the field of human reproduction and embryology, numerous groups are participating in the development of AI algorithms for ART with significant variability in the expertise and resulting published research. These ART/AI publications should enforce common minimum standards to permit fully transparent peer-review of the studies. These include utilizing sufficient amounts of high-quality data and proper development paths for AI algorithms, in full compliance with existent international regulations. Notably, what these minimum criteria are difficult to define and will depend on the dataset utilized.

Clinically relevant endpoints for AI in fertility

The potential of AI opportunities to investigate in fertility is very broad and largely extends through the entire IVF patient’s journey (Fig. 1).

Table 1 Peer-reviewed publications of AI algorithms compared with doctors [16]

Specialty	Images	Publication
Radiology/neurology	CT head, acute neurological events	Titano et al. [17]
	CT head for brain hemorrhage	Arbabshirani et al. [18]
	CT head for trauma	Chilamkurthy et al. [19]
	CXR for metastatic lung nodules	Nam et al. [20]
	CXR for multiple findings	Singh et al. [21]
	Mammography for breast density	Lehman et al. [22]
	Wrist X-ray	Lindsey et al. [23]
Pathology	Breast cancer	Ehteshami Bejnordi et al. [24]
	Lung cancer (+ driver mutation)	Coudray et al. [25]
	Brain tumors (+ methylation)	Capper et al. [26]
	Breast cancer metastases	Steiner et al. [27]
	Breast cancer metastases	Liu et al. [28]
Dermatology	Skin cancers	Esteva et al. [29]
	Melanoma	Haenssle et al. [30]
	Skin lesions	Han et al. [31]
Ophthalmology	Diabetic retinopathy	Gulshan et al. [32]
	Diabetic retinopathy	Abramoff et al. [33]
	Diabetic retinopathy	Kanagasingam et al. [34]
	Congenital cataracts	Long et al. [35]
	Retinal diseases (OCT)	De Fauw et al. [36]
	Macular degeneration	Burlina et al. [37]
	Retinopathy of prematurity	Brown et al. [38]
	AMD and diabetic retinopathy	Kermany et al. [39]
Gastroenterology	Polyps at colonoscopy	Mori et al. [40]
	Polyps at colonoscopy	Wang et al. [41]
Cardiology	Echocardiography	Madani et al. [42]
	Echocardiography	Zhang et al. [43]

Data-driven solutions might help detect early indicators of infertility allowing an earlier identification of potential patients to refer for treatment as well as support doctors during a consultation to define a precise treatment strategy, providing estimations of success in terms of pregnancy and take-home baby rates.

Before starting treatment, we may also be able to define the best protocol of stimulation to apply for each patient to maximize success. We can also define the need of adjusting drugs and doses, during treatment, based on real-time biomarkers of follicle development. In the embryology and andrology laboratories, we could use data-driven solutions to evaluate gamete viability, embryo implantation potential, and uterine receptivity status before embryo transfer, therefore allowing us to target defined windows of implantation potential. Finally, we may be able to provide individualized support for the luteal phase. All this would allow the improvement of pregnancy while reducing the time to achieve a healthy live birth.

In summary, widely accepted clinically relevant endpoints to target are as follows: (1) the birth of a single and healthy baby, (2) decrease of miscarriages, (3) higher implantation and clinical and ongoing pregnancy rates, (4) objective assessments of cytoplasmic and genetic health of gametes and embryos, and (5) standardization of operator assessments.

The quality and quantity of data is key

The availability of data has increased in the last decades, and the rapid developments and advances of machine learning have sparked enthusiasm for applying data science techniques to solve problems and improve current status in clinical care [45]. However, it often happens that only retrospective data is available, leading to a very expensive post-process for sorting and labeling/annotating the data (if possible) or making it necessary to collect new data. Labels are indeed the most

Table 2 Examples of some AI technologies related to ART

Target	Objective	Company
Embryo assessment	EEVA viability status	https://hcp.merckgroup.com/en/fertility/technologies
	Geri Assess 2.0	https://hcp.merckgroup.com/en/fertility/technologies
	Viability and implantation potential	www.Presagen.com
	Viability and implantation potential	www.IVIVision.ai
	Viability and implantation potential	www.FutureFertility.com
	Viability and implantation potential	www.Binflux.com
	Viability and implantation potential	www.IVY.ai
	Viability and implantation potential	STORK
	Viability and implantation potential	Fairtility
	Euploidy/aneuploidy status	www.Presagen.com
	Euploidy/aneuploidy status	ERICA— www.embryoranking.com
	Euploidy/aneuploidy status	VIOLET— www.FutureFertility.com
Oocyte assessment	Viability status	www.FutureFertility.com
Stimulation protocol	Predict live birth	www.UNIVFY.com
	Predict pregnancy	www.apricity.life
Uterine receptivity	Predict receptivity potential	www.FutureFertility.com

valuable part of data for any supervised machine learning model. Data quality and data quantity are challenges that need to be overcome to develop reliable and robust algorithms [46]. The amount of available data is particularly relevant in ART, given the small general population of patients being treated. Given a problem, the information content of the collected data is relevant as a model can only learn from the data that is used for its training. The complexity of the problem to solve and the complexity of the model itself are influencing factors for data sufficiency. Domain experts can help to estimate how representative the collected data is for a given defined problem. Additionally, measures from information theory (e.g., entropy, mutual information, and information gain) can be used to quantify the amount of information present in the data and guide data collection and data modeling, assessing, for example, the introduction of new variables or the collection of new samples [47]. Moreover, data size requirements might depend on the problem dimensionality: a higher number of features likely lead to a larger amount of data needed for analysis and modeling [48]. In the case of a high number of variables, the search problem space can be large, while available data might be sparse with potentially redundant information. To evaluate and quantify the needed sample size while training a model, learning curves can be used, describing the model performance as a function of the training size. While the number of training samples increases, the model performance increases rapidly at the beginning, subsequently slowing down until a point where the model performance improves only

marginal with an increasing number of samples. Several studies have been published to characterize performances and generalization error with data size [49, 50].

An additional challenge to consider may be the heterogeneity of data sources generated by different types of equipment and not centralized data collection systems. Every ART clinic has implemented different workflows and standard procedures that need to be considered during data analysis and modeling. The standardization of the processes generating data is a promising and important effort that all involved organizations should likely embrace. A definition and consequent implementation of quality standard procedures would facilitate the modeling of data coming from different sources reducing data variability. This would also help in a broad validation of any algorithm before potential production and commercialization. Data collection is the first imperative step, but it requires coming with quality and quantity, depending on the problem in question. Moreover, it is crucial to develop and implement proper modeling approaches to better understand the data, extract information, and make reliable generalizations under proper assumptions and limitations.

Proper development path of AI algorithms for fertility

The development of AI-based solutions for medical decisions in many therapeutic areas is challenging and might be even

Fig. 1 Summary scheme of how AI can improve IVF treatments throughout the process

more complex for ART for several reasons including (1) small general population of patients being treated, leading to small data sets, (2) AI solutions infertility are complicated to patent and protect, (3) lack of standardization leads to no homogeneous data sets, (4) data is stored in diverse ecosystems, (5) data protection laws and regulations differ by country, (6) no technical ecosystem that allows applying AI solutions at scale, and (7) there are unclear regulatory approval routes that differ by country. In the treatment of fertility, data generation is fragmented and makes the application of AI solutions a challenge.

We must also take into consideration that many healthcare providers do not fully understand the technology and therefore shy away from its applicable relevance not supporting its development. Indeed, the complexity of fertility procedures applied by different doctors and embryologists and the need for continuous training to secure the collection of high-quality data to be used for the development of any AI solution can make it difficult to understand and accept.

To build a successful decision support system, several factors need to be considered. A minimal potential workflow to address and model a clinically relevant problem can be summarized as follows: (1) identify and define clinically a relevant problem, (2) design the study including patients and clinicians, (3) collect a large amount of quality data, automatically if possible, in an ecosystem that allows collection and storage of entire cycle data (electronic medical record), (4) label or annotate the data, (5) correct for biases and errors, (6) develop model/s to solve the considered problem rigorously following

ML principles and best practice rules to avoid overfitting and to obtain a robust solution, (7) validate the selected solution, and (8) monitor and transparently provide oversight and accountability of the methodology applied for training, test, and validation. Depending on the scope of the solution and the possibility to commercialize the technology, different levels of validation will be required. Additionally, a more systematic recording process should be implemented to automatize tasks such as identify unusual or outlier laboratory results. A robust predictive algorithm that provides probability estimations needs to be trained, tested, and properly validated in randomized clinical trials, to support clinically relevant data-driven decisions. Eventually, it is important to consider potential ethical issues and develop a friendly user interface for healthcare providers to facilitate the use of such a system.

Regulations related to the development of medical software solutions

Medical software allows healthcare operators to access clinical information stored in electronic health records and additionally may support clinical decisions. These systems must be designed, built, validated, and maintained following strict global regulations and standards to ensure reliability and safety. The International Electrotechnical Commission specifies the life cycle requirements for the development of medical software and software within medical devices, harmonized by the EU/USA. It describes a set of processes, activities,

and tasks, establishing a common framework for the life cycle processes of medical device software. Thus, it should be taken into consideration that the development of any AI-related technologies for fertility must follow the International Standard IEC 62304 (<https://www.iso.org/obp/ui/#iso:std:iec:62304:ed-1:v1:en>).

The Organization for Economic Cooperation and Development (OECD), a unique forum where the governments of 36 member states with market economies work with each other, as well as with more than 70 non-member economies to promote economic growth, prosperity, and sustainable development, defined AI principles that must be supported and in particular the need to “commit to transparency and responsible disclosure” in the use of AI systems. Essential points to consider should be the following: (1) requirement of disclosure to provide public assurances of liability taken by different organizations that are developing AI; (2) existence of rules for evaluating the level of potential damage to the technology; (3) ensure transparency, without revealing source code, trade secrets, or intellectual property; (4) any AI system on the market that is making determinations or recommendations with potentially significant implications for people should be able to explain and contextualize how and why it reached a particular conclusion; and (5) it is necessary to test AI for biases, carrying out test for fairness, bias, robustness, and security and take corrective measures if necessary [51].

Conclusions

We must recognize the importance of continually innovating in the fertility space to improve outcomes for the benefit of patients. Objectivity must be brought to the decision-making process in every step of ART to help more precisely determine the right treatment for the circumstance. This includes assessing the cause of infertility; determining type and decision points of the stimulation protocol; and accurately assessing gametes viability, embryo implantation, and uterine receptivity potential. The use of algorithms/AI to help guide human decision-making in each of these critical areas in ART has the potential to reduce subjectivity and improve consistency. However, it is necessary to be cautious and ensure that appropriate approaches are followed in developing accurate predictive algorithm models.

Another potential issue with universal use of AI in treatment of infertility arises from clinic competition and intellectual property. Will developed algorithms/AI become a commodity, specific for a particular clinic, or will these decision-making tools be widely available and applicable between centers? Certainly the hurdles that accompany data confidentiality and clinic competition will need to be addressed.

In summary, fertility relevant areas where AI applications could be focusing in the medium or short term should be the treatment-related areas like diagnostics; the selection of precise drugs and doses as well as adjustment during treatment; the homogeneity of the evaluation by the operators; and assessment of genetic quality, viability of blastocysts, implantation, and clinical and ongoing pregnancy rates. A very fundamental aspect to build any predictive algorithm is data, in a sufficient quantity and quality. Although the minimum data size requirements might be difficult to determine as it depends on the final point to consider, the amount of data required for training, test, and validation of an algorithm to reach a valuable and accurate clinical support is critical. A precise definition of quality standards would help in a potential reduction of variability and promote collaborations and projects. It is necessary to secure that any trained, validated, and robust models follow IEC 62304 regulatory requirements. Finally, it is important to re-emphasize that if our intention is to utilize AI-based algorithms and to potentially improve precision and outcomes for the benefit of patients, we must start by securing not only those algorithms that were based on proper paths for development and validation but also on a foundation of big quality data ecosystems.

Acknowledgments Fertility AI Forum Group: Gerard Letterie, Integramed; Pascual Sánchez, Ginemed; Geoff Trew, The Fertility Partnership; Jason Swain, CCRM Management Co.; Marcos Meseguer, IVIRMA; Dan Nayot, Trio Fertility; Alison Campbell, CARE; Ian Huang, Storck-Binflux; Jan Choma, Cognexa; Kevin Loewke, DANA; Maria Paola Piqueras, Ginemed; Paul Nader, Baby Sentry; Michael Schindler, Meditex; Marck Marcom, Ideas EMR; Ed Vom, Planet Innovation; Eleanora Lippolis, Merck; Sebastian Bohl, Merck, Jan Kirsten, Merck; Daniel Abshagen, Merck; Diego Ezcurra, Merck.

References

1. European IVF-Monitoring Consortium (EIM) for the European Society of Human Reproduction and Embryology (ESHRE), Calhaz-Jorge C, et al. Assisted reproductive technology in Europe, 2012: results generated from European registers by ESHRE. *Hum Reprod*. 2016;31:1638–52.
2. de Mouzon J, Goossens V, Bhattacharya S, Castilla JA, Ferraretti AP, Korsak V, et al. Assisted reproductive technology in Europe, 2006: results generated from European registers by ESHRE. *Hum Reprod*. 2010;25:1851–62.
3. Adamson GD, de Mouzon J, Chambers GM, Zegers-Hochschild F, Mansour R, Ishihara O, et al. International committee for monitoring assisted reproductive technology: world report on assisted reproductive technology, 2011. *Fertil Steril*. 2018;110:1067–80.
4. Centres for Disease Control and Prevention. Assisted reproductive technology: ART trends 2002–2011. <http://www.cdc.gov/art/ART2011/section5.htm#f43>. Accessed 16 Aug 2016.
5. Curchoe CL, Bormann CL. Artificial intelligence and machine learning for human reproduction and embryology presented at ASRM and ESHRE 2018. *J Assist Reprod Genet*. 2019;36:591–600.

6. Letterie GS, MacDonald A. A computerized decision–support system for day to day management of ovarian stimulation cycles during in vitro fertilization. *Fertil Steril*. 2019;112:e28.
7. Khosravi P, et al. Robust automated assessment of human blastocyst quality using deep learning. *bioRxiv*. 2018;394882. <https://doi.org/10.1101/394882>.
8. Hoo-Chang S, et al. Deep convolutional neural networks for computer-aided detection: CNN architectures, dataset characteristics and transfer learning. *IEEE Trans Med Imag*. 2016;35:1285–98.
9. Turing AM. On computable numbers, with an application to the entscheidungsproblem. a correction. *Proc Lond Math Soc*. 1938;2:43:544–6.
10. Turing AM. Computing machinery and intelligence. *Mind*. 1950;236:433–60.
11. McCulloch W, Pitts W. A logical calculus of the ideas immanent in nervous activity (reprinted from 1943). *Bull Math Biol*. 1990;52:99–115.
12. Krizhevsky A, Sutskever I, Hinton G. ImageNet classification with deep convolutional neural networks. *Handb. Approx. Algorithms Metaheuristics*. 2007;45:1–45–16. <https://doi.org/10.1201/9781420010749>.
13. The AI effect. How artificial intelligence is making health care more human. MIT Technology reviews insights. 2019.
14. Tran D, Cooke S, Illingworth PJ, Gardner DK. Deep learning as a predictive tool for fetal heart pregnancy following time-lapse incubation and blastocyst transfer. *Hum Reprod*. 2019;34:1011–8.
15. Shen D, Wu G, Suk H. Deep learning in medical image analysis. *Physiol Behav*. 2017;176:139–48.
16. Topol EJ. High-performance medicine: the convergence of human and artificial intelligence. *Nat Med*. 2019;25:44–56.
17. Titano JJ, Badgeley M, Schefflein J, Pain M, Su A, Cai M, et al. Automated deep-neural-network surveillance of cranial images for acute neurologic events. *Nat Med*. 2018;24:1337–41.
18. Arbabshirani MR, Fornwalt BK, Mongelluzzo GJ, Suever JD, Geise BD, Patel AA, et al. Advanced machine learning in action: identification of intracranial hemorrhage on computed tomography scans of the head with clinical work flow integration. *NPJ Digit Med*. 2017;1:9. <https://doi.org/10.1038/s41746-017-0015-z>.
19. Chilamkurthy S, et al. Articles Deep learning algorithms for detection of critical findings in head CT scans: a retrospective study. *Lancet*. 2018;6736:1–9.
20. Nam JG, Park S, Hwang EJ, Lee JH. Development and validation of deep learning – based automatic detection algorithm for malignant pulmonary nodules on chest radiographs. *Radiology*. 2019;290:218–28.
21. Singh R, et al. Deep learning in chest radiography: detection of findings and presence of change. *PLoS ONE*. 2018;13(13):1–12.
22. Lehman CD, et al. Mammographic breast density assessment using deep learning: clinical implementation. *Radiology*. 2018;00:1–7.
23. 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 USA*. 2018;115:11591–6.
24. Bejnordi BE, et al. Diagnostic assessment of deep learning algorithms for detection of lymph node metastases in women with breast cancer. *JAMA*. 2017;318:2199–210.
25. Coudray N, et al. images using deep learning. *Nat Med*. 2018;24:1559–69.
26. Capper D, Jones DTW, Sill M, Hovestadt V, Schrimpf D, Sturm D, et al. DNA methylation-based classification of central nervous system tumours. *Nature*. 2018;555:469–74.
27. Steiner DF, MacDonald R, Liu Y, Truszowski P, Hipp JD, Gammage C, et al. Impact of deep learning assistance on the histopathologic review of lymph nodes for metastatic breast cancer. *Am J Surg Pathol*. 2018;42:1636–46.
28. Liu Y, Kohlberger T, Norouzi M, Dahl GE, Smith JL, Mohtashami A, et al. Artificial intelligence – based breast cancer Nodal. *Arch Pathol Lab Med*. 2019;143:859–68.
29. Esteva A, et al. Dermatologist-level classification of skin cancer with deep neural networks. *Nat Publ Group*. 2017;542:115–8.
30. Haenssle HA, Fink C, Schneiderbauer R, Toberer F, Buhl T, Blum A, et al. Man against machine: diagnostic performance of a deep learning convolutional neural network for dermoscopic melanoma recognition in comparison to 58 dermatologists. *Ann Oncol*. 2018;29:1836–42.
31. Han SS, Kim MS, Lim W, Park GH, Park I. Classification of the clinical images for benign and malignant cutaneous tumors using a deep learning algorithm. *J Invest Dermatol*. 2018;138:1529–38.
32. Gulshan V, et al. Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. *JAMA*. 2016;94043:1–9.
33. Abramoff MD, Lavin PT, Birch M, Shah N, Folk JC. Pivotal trial of an autonomous AI-based diagnostic system for detection of diabetic retinopathy in primary care offices. *NPJ Digit Med*. 2018. <https://doi.org/10.1038/s41746-018-0040-6>.
34. Kanagasigam Y, Xiao D, Vignarajan J, Preetham A, Mehrotra A. Evaluation of artificial intelligence – based grading of diabetic retinopathy in primary care. *JAMA*. 2018;1:1–6.
35. Long E, et al. An artificial intelligence platform for the multihospital collaborative management of congenital cataracts. *Nat Biomed Eng*. 2017;0024:1–8.
36. De Fauw J, et al. Clinically applicable deep learning for diagnosis and referral in retinal disease. *Nat Med*. 2018;24:1342–54.
37. Burlina PM, Joshi N, Pekala M, Pacheco KD, Freund DE, Bressler NM. Automated grading of age-related macular degeneration from color fundus images using deep convolutional neural networks. *JAMA Ophthalmol*. 2017;135:1170–6.
38. Brown JM, Campbell JP, Beers A, Chang K, Ostmo S, Chan RVP, et al. Automated diagnosis of plus disease in retinopathy of prematurity using deep convolutional neural networks. *JAMA Ophthalmol*. 2018;136:803–10.
39. Kermany DS, Goldbaum M, Cai W, Lewis MA. Identifying medical diagnoses and treatable diseases by image-based deep learning resource. *Cell*. 2018;172:1122–1131.e1129.
40. Mori Y, Kudo SE, Misawa M, Saito Y, Ikematsu H, Hotta K, et al. Real-time use of artificial intelligence in identification of diminutive polyps during colonoscopy. *Ann Intern Med*. 2018;169:357–66.
41. Wang P, Xiao X, Glissen Brown JR, Berzin TM, Tu M, Xiong F, et al. Development and validation of a deep-learning algorithm for the detection of polyps during colonoscopy. *Nat Biomed Eng*. 2018;2:741–8.
42. Madani A, Arnaout R, Mofrad M. Fast and accurate view classification of echocardiograms using deep learning. *NPJ Digit Med*. 2018;1–8. <https://doi.org/10.1038/s41746-017-0013-1>.
43. Zhang J, Gajjala S, Agrawal P, Tison GH, Hallock LA, Beussink-Nelson L, et al. Fully automated echocardiogram interpretation in clinical practice feasibility and diagnostic accuracy. *Circulation*. 2018;138:1623–35.
44. Fujisawa Y, Otomo Y, Ogata Y, Nakamura Y, Fujita R, Ishitsuka Y, et al. Deep-learning-based, computer-aided classifier developed with a small dataset of clinical images surpasses board-certified dermatologists in skin tumour diagnosis. *Br J Dermatol*. 2019;180:373–81.
45. Celi LA, Csete M, Stone D. Optimal data systems: the future of clinical predictions and decision support. *Curr Opin Crit Care*. 2014;20:573–80.
46. Hee, K. Is data quality enough for a clinical decision?: apply machine learning and avoid bias. *Proc. - 2017 IEEE Int. Conf. Big Data, Big Data 2017 2018-Janua*, 2612–2619. 2017.

47. MacKay DJC. Information theory, inference, and learning algorithms. Cambridge University Press 2003; 2005. <https://doi.org/10.1166/asl.2012.3830>.
48. Bellman RE. Dynamic Programming; Princeton University Press; 2010.
49. Cho J et al. How much data is needed to train a medical image deep learning system to achieve necessary high accuracy. Conf Pap ICLR 2016 HOW. 2016.
50. Hestness J et al. Deep learning scaling is predictable, empirically. arXiv:1712.00409. 2017.
51. Hagemann BR, Leclerc J. Precision regulation for artificial intelligence. IBM Policy Lab 1–5.

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