2021

Interpretable, not black-box, artificial intelligence should be used for embryo selection

Michael Anis Mihdi Afnan

Yanhe Liu
Edith Cowan University

Vincent Conitzer

Cynthia Rudin

Abhishek Mishra

See next page for additional authors

10.1093/hropen/hoab040
Afnan, M. A. M., Liu, Y., Conitzer, V., Rudin, C., Mishra, A., Savulescu, J., & Afnan, M. (2021). Interpretable, not black-box, artificial intelligence should be used for embryo selection. Human reproduction open, 2021(4), hoab040. https://doi.org/10.1093/hropen/hoab040

Human Reproduction Open, pp. 1, 2021
https://doi.org/10.1093/hropen/hoab040

human reproduction open

OPINION

Interpretable, not black-box, artificial intelligence should be used for embryo selection

Michael Anis Mihdi Afnan [1,*], Yanhe Liu[2,3,4,5], Vincent Conitzer[6,7,8,9,10], Cynthia Rudin[6,11,12], Abhishek Mishra[13], Julian Savulescu [13,14,15], and Masoud Afnan[16]

[1]Wrightington, Wigan and Leigh NHS Foundation Trust, Greater Manchester, UK [2]Monash IVF Group, Southport, Australia [3]School of Human Sciences, University of Western Australia, Crawley, Australia [4]School of Medical and Health Sciences, Edith Cowan University, Joondalup, Australia [5]School of Health Sciences and Medicine, Bond University, Robina, Australia [6]Department of Computer Science, Duke University, Durham, NC, USA [7]Department of Economics, Duke University, Durham, NC, USA [8]Department of Philosophy, Duke University, Durham, NC, USA [9]Department of Computer Science, Institute for Ethics in AI, University of Oxford, Oxford, UK [10]Department of Philosophy, Institute for Ethics in AI, University of Oxford, Oxford, UK [11]Department of Electrical Engineering, Duke University, Durham, NC, USA [12]Department of Statistical Science, Duke University, Durham, NC, USA [13]Uehiro Centre for Practical Ethics, University of Oxford, Oxford, UK [14]Wellcome Centre for Ethics and Humanities, University of Oxford, Oxford, UK [15]Murdoch Children's Research Institute, Royal Children's Hospital, Parkville, Australia [16]Department of Obstetrics & Gynaecology, Qingdao United Family Hospital, Qingdao, China

*Correspondence address. Royal Albert Edward Infirmary, Wigan Lane, Wigan WN1 2NN, UK. E-mail: michaelafnan@icloud.com
https://orcid.org/0000-0001-5156-7806

Submitted on August 02, 2021; resubmitted on October 18, 2021; editorial decision on October 25, 2021

Abstract: Artificial intelligence (AI) techniques are starting to be used in IVF, in particular for selecting which embryos to transfer to the woman. AI has the potential to process complex data sets, to be better at identifying subtle but important patterns, and to be more objective than humans when evaluating embryos. However, a current review of the literature shows much work is still needed before AI can be ethically implemented for this purpose. No randomized controlled trials (RCTs) have been published, and the efficacy studies which exist demonstrate that algorithms can broadly differentiate well between 'good-' and 'poor-' quality embryos but not necessarily between embryos of similar quality, which is the actual clinical need. Almost universally, the AI models were opaque ('black-box') in that at least some part of the process was uninterpretable. This gives rise to a number of epistemic and ethical concerns, including problems with trust, the possibility of using algorithms that generalize poorly to different populations, adverse economic implications for IVF clinics, potential misrepresentation of patient values, broader societal implications, a responsibility gap in the case of poor selection choices and introduction of a more paternalistic decision-making process. Use of interpretable models, which are constrained so that a human can easily understand and explain them, could overcome these concerns. The contribution of AI to IVF is potentially significant, but we recommend that AI models used in this field should be interpretable, and rigorously evaluated with RCTs before implementation. We also recommend long-term follow-up of children born after AI for embryo selection, regulatory oversight for implementation, and public availability of data and code to enable research teams to independently reproduce and validate existing models.

Key words: IVF / embryo selection / artificial intelligence / AI / machine learning / ML / interpretable / black-box / ethics

Introduction

Competitive embryo selection is unique to IVF. Traditional embryo selection is based on several snapshot observations of an embryo under a microscope, at specific time points during culture. Considering the dynamic nature of embryo development, the static nature of the information collected in this method limits the accuracy of embryo selection (Gardner *et al.*, 2015). Examples of techniques developed to select embryos more likely to implant include: extended embryo culture (Gardner *et al.*, 2000); time-lapse imaging of the embryo in culture over a number of days (Liu *et al.*, 2016); metabolomic profiling of spent culture media (Zmuidinaite *et al.*, 2021); and pre-implantation genetic testing (Kemper *et al.*, 2020). Each of these techniques potentially give rise to many thousands of data points.

Furthermore, the evaluation of embryo quality by the embryologist is limited by considerable inter-operator variability, due to the current

mix of objective and subjective measures in assessment, and human factors, such as being influenced by confounders (Liu *et al.*, 2019).

Embryo freezing techniques have improved dramatically such that if all the embryos are transferred, albeit one at a time, embryo selection would not affect pregnancy or live birth rates per *egg collection*, as ultimately all embryos will be given the chance to implant. However, not all couples persist with treatment even in the presence of remaining frozen embryos (Centers for Disease Control and Prevention, US Department of Health and Human Services, 2018; HFEA 2020). For some couples, therefore, maximizing their chance of a live birth at an earlier transfer could raise their overall chances of having a baby.

The rapid development of artificial intelligence (AI) in recent years has made possible the ability to objectively process and interpret vast quantities of data, both visual and tabular, to potentially improve embryo selection. Whilst we acknowledge that current methods of embryo selection have limitations and risks, we are concerned that these might be exacerbated by AI, and there may be new ones faced.

For a further exploration of the issues raised in this opinion paper, but aimed at a computer science and ethics audience, we refer the reader to a paper presented by the same authors at the AI, Ethics and Society 2021 conference (Afnan *et al.*, 2021).

Advantages of using AI for embryo selection

Given the known desired outcome of a healthy live birth, AI and machine learning (ML) in particular, is ideally placed to compute and make sense of complex data, to look for subtle patterns unobserved by embryologists (Fernandez *et al.*, 2020). AI can standardize and automate many of the processes, reducing inter-observer variability and providing a more objective embryo assessment (Rosenwaks, 2020).

ML gives rise to models that can automatically learn and adapt as they are exposed to more data (whether images or other data). This is particularly useful when there is access to lots of data, but we do not immediately know how to leverage it to make better predictions, or when we cannot manually process it all to generate meaningful knowledge. Computer vision (CV) allows large amounts of image data to be automatically analyzed by algorithms, and rapid recent advances offer promise to improve embryo selection.

The most popular ML models for CV are uninterpretable ('black-box') models. These are either too complicated for any human to understand, or they are proprietary—in which case, comprehension of such a model is not possible for outsiders (Rudin, 2019). In contrast, an interpretable ML model is a predictive model that is constrained so that a human can better understand its reasoning process (Rudin, 2019; Rudin *et al.*, 2021). As we will show in this work, interpretability has important epistemic and ethical implications.

Current use of AI in embryo selection

We searched MEDLINE, Embase and Google Scholar up to and including February 2021 for full-text studies evaluating AI to select embryos using the strategy included in the Supplementary Data. We checked the citations of papers we identified in the search for any publications we might have missed.

Studies evaluating AI for embryo selection make impressive accuracy claims for their ML models (Khosravi *et al.*, 2019; Tran *et al.*, 2019). One commonly reported performance measure is the receiver operating characteristic (ROC) curve which shows how a test's sensitivity and specificity correlate at different thresholds. The AUC indicates the test's performance. An AUC >0.9 usually indicates outstanding performance, and the ML models from the studies cited above surpass this benchmark.

Studies that evaluate the efficacy of AI models for embryo selection do so for two types of outcomes: first, outcomes meaningful to the patient, such as a live birth or a fetal heartbeat (FH) positive pregnancy; or, second, agreement with the existing standard, which in this case would be assessment by embryologists. One of the challenges of using live birth as the meaningful outcome (ground truth) is that a potentially viable embryo can result in either a live birth, or no live birth, depending on other, non-embryo factors, such as the health of the mother.

Tran *et al.*'s (2019) study belongs to the first category. They evaluated a model called IVY, which rates how likely an embryo is to lead to an FH pregnancy on a confidence scale of 0 (definitely will not implant) to 1 (definitely will implant). Their ROC curve's AUC was 0.93. However, as Kan-Tor *et al.* (2020) point out, the majority of the embryos on which the algorithm had been trained and tested were of such poor quality that they would have been discarded in any event, thereby artificially inflating the AUC. As Kan-Tor *et al.* explain, the clinical need is to identify the embryo with the highest chance of success among a set of embryos that appear to be potentially viable, and not from embryos which embryologists readily discard.

Khosravi *et al.*'s (2019) study, on the other hand, belongs to the second category. They categorized embryos into three groups—good-, fair- and poor-quality embryos according to a consensus of multiple embryologists. They then evaluated their AI algorithm's ability to identify the good- and the poor-quality embryos (but not the fair-quality embryos); for this task, the algorithm achieved 96.94% accuracy. This was better than the performance of individual embryologists. However, broad categorizations into 'good' or 'poor' quality are of limited benefit when trying to find the best embryo in a group of similar-quality embryos.

The earlier analyses of the Khosravi *et al.* (2019) and Tran *et al.* (2019) studies demonstrate the importance of understanding exactly how researchers test their algorithms before drawing conclusions from headline statistics. These studies are important steps to investigate efficacy (the ability to produce a specified outcome in experimental circumstances), to develop the tool and establish proof of principle. However, they are only a prelude to testing in the clinic. When Curchoe *et al.* (2020) reviewed how the results of AI studies in reproductive medicine relate to real-life clinical practice, they highlighted four pitfalls that are common throughout the literature: small sample sizes, imbalanced datasets, non-generalizable settings and limited performance metrics.

We also point out that many studies in this field use neural networks that are not interpretable, and not designed to be interpretable, despite the existence of literature on interpretable neural networks that reports achieving comparable accuracy to black-box neural networks (e.g., Chen *et al.*, 2019). Other approaches use interpretable features (whether they are labeled manually by doctors or labeled by neural networks whose output can be manually verified) but combine them in uninterpretable ways, such as using principal component analysis pre-processing (which forces a dependence on all variables) followed by an ML method such

as a neural network or random forest (Milewski *et al.*, 2017; Chavez-Badiola *et al.*, 2020). The work of Leahy *et al.* (2020) is more interpretable because the model is decomposable into separate neural network models that each extract different information (e.g., measurements of an embryo) that can be directly checked by an embryologist. Raef *et al.* (2020) and Morales *et al.* (2008) created interpretable hand-calculated features and applied a variety of classical ML algorithms to them. However, the majority of studies feature uninterpretable (or 'black-box') AI algorithms which have a number of pitfalls.

Epistemic problems of black-box AI models

Black-box models create information asymmetries

The use of black-box models creates an information asymmetry between the company selling the tool and the clinicians having to make daily decisions as to which embryo to transfer. Such models would force the embryologist to abrogate decision-making to programs they do not understand. It is not possible to fully evaluate whether to trust these complex models without an understanding of their reasoning processes. For example, black-box models have been introduced in radiology, a quintessential example of CV applied to medical imaging, and which have been tested sufficiently to gain US Food and Drug Administration approval for clinical practice. Unfortunately, these have not done well, and no-one knows why (Voter *et al.*, 2021), raising concerns about their generalizability (O'Connor, 2021).

Confounders are rampant

One of the reasons why black-box systems do not generalize despite succeeding in clinical trials, and therefore receiving regulatory approval, is confounding. If we do not understand what a black-box model is doing, it is entirely possible that its predictions are based on confounders that should not be used as predictors, such as age. Confounders are often difficult to detect and cause models not to generalize. When coupled with a poor choice for evaluation metric, the confounding might not be noticed (O'Connor, 2021; Voter *et al.*, 2021).

Perpetuated bias

ML-assisted embryo selection is an ongoing process whereby the machine learns and adapts its algorithms from the previous generation of data. Should a bias occur at any point, then not only would that bias be perpetuated, but it would also be magnified in future generations as the machine will learn from biased data. It may be argued that such bias may occur as a result of embryologist embryo selection. However, ML models could systematize and embed the bias in the selection process to a greater extent than embryologists.

Real-time error-checking is harder with black-box models

The two problems discussed above (information asymmetry and the possibility of confounders) lead to a third problem: the difficulty of error-checking the model in real-time as it makes predictions in the clinic. We would want the clinician to be able to determine whether the model is reasoning in a way that is obviously wrong and catch new problems immediately should they arise.

The economics of 'buying into' a brittle model does not favor clinicians or patients

A potential consequence of the problems of information asymmetry and confounders listed above would be that black-box model performance may be brittle to changes from the system it was trained on, and thus would likely be limited to the ecosystem in which it has been shown to work. This means that a clinic using this model may need to buy into that ecosystem, ovarian stimulation regimens, use of incubators and culture medium amongst other potential variables. This gives AI companies a great deal of economic power over clinics, potentially increasing treatment costs. This could be mitigated by models that are robust across domains. Interpretable AI is easier to make robust because it is known how it works.

Overall troubleshooting is difficult for black-box models

If the model were more interpretable, it might be easier to troubleshoot broad problems (beyond serious issues that might be noticed in real-time usage). This includes ethical concerns, such as sex, disability or racial bias (which we will discuss), as well as epistemic issues with accuracy or subtle confounding.

Black-box models are difficult to 'explain'

There is a growing body of work on 'explaining' black-box models. However, such explanations are problematic for reasons outlined by Rudin (2019). For instance, explanations for black-box models are often not faithful (e.g., claiming that race is used in a model when instead, a factor correlated with race was used), explanations for visual images tend to be incomplete (highlighting parts of the image without explaining how these parts of the image are used to make a prediction), and different explanation methods can produce completely different explanations. Interpretable models are different because they are self-explanatory. Explaining a black-box model lends unwarranted authority to it (Rudin and Radin, 2019), which might deter development of an inherently interpretable model with the same accuracy.

Ethical concerns with black-box AI models

Compromised shared decision-making

Over the past few decades, clinical practice has shifted from a paternalistic model to a model of shared decision-making aimed at promoting patient autonomy (Charles *et al.*, 1997). Opaque AI models compromise shared decision-making due to the inability of the clinician and patient to understand the model's decision (Bjerring and Busch, 2020), for example, information as to why a particular embryo is selected or is not selected (such as, the number and symmetry of the cells, or if the cells are fragmented, and therefore what the chances of implantation are, and why implantation may fail).

There are counterarguments to this concern, which should also be considered (Mishra et al., In press). It will be important to fully explain what is known about how the AI model comes to a 'decision' (nature and size of dataset, reasons for confidence in prediction, possible alternative lines of justification, etc.), and further examine how interactions between clinicians and patients may change (see below). Existing measures of shared decision-making and decision quality, such as the Decision Conflict Scale (Garvelink et al., 2019), the OPTION Scale (Elwyn et al., 2003) and the SURE Test (Légaré et al., 2010) (among other patient-reported measures) can be used.

It is important, however, to contextualize this concern. Firstly, current expert judgment is not very accurate at predicting a live birth. Secondly, autonomy requires understanding information relevant and meaningful to one's values. Knowing the basis of a prediction (cleavage rate, symmetry, etc.) is not relevant: what is relevant are the risks, side-effects and benefits, and the confidence attached to these assessments.

However, black-box AI has the potential to significantly undermine shared decision-making in a way that interpretable AI does not. Although they are marketed and approved as decision aids to the clinician, where the decision finally rests with the clinicians, black-box AI will in practice have the potential to introduce a new form of paternalism: 'machine paternalism.' It is known that people tend to be complacent about use of automation (Parasuraman and Manzey, 2010) and accepting AI in a role when they are familiar with AI in that role or when they believe it performs well (Kramer et al., 2018). In practice, it is hard to see how clinicians will challenge the deliverances of black-box AI. Indeed, doing so without good reason might open them up to legal liability. So, in practice, black-box AI risks instrumentalizing clinicians and replacing human decision-making.

Misrepresentation of patient values

AI may misrepresent patient values. For example, there are reported differences between early morphokinetic profiles of male and female embryos (Tarín et al., 2014; Bronet et al., 2015; Wang et al., 2018; Huang et al., 2019) (and other traits might be similarly differentially represented at this early stage). Models for embryo selection run the risk of systematically selecting for these traits if they are perceived by the model to be correlated with implantation success. For example, if a patient prefers that sex be randomly selected, this model may run counter to those values. If such models are opaque, this systematic favoring of particular traits may not be detected and so cannot be corrected for in a way that it could be with interpretable models.

Again, it is important not to overstate this concern. There have been calls for systems to be 'value flexible' (McDougall, 2019). The patient's own values could be inserted into AI algorithms (e.g., preference for sex and other non-disease characteristics) and AI might bring to the surface the importance of these values in decision-making. Of course, valuing and selecting non-disease traits (such as sex or intelligence) raises the debate around designer babies, but some have argued that such selection is permissible (Agar, 2004) or even a moral obligation when it relates to the well-being of a future child (Savulescu, 2001; Savulescu and Kahane, 2009, 2016).

Health and well-being of future children

Such potential biasing of AI-selection might also have impacts on the health or well-being of future children. For example, it is possible that some disadvantageous trait (such as increased risk of cancer or mental disorder) correlates with a higher chance of implantation. This also underscores the importance of clinical trials not merely measuring implantation or even healthy live birth but long-term well-being of the child created by IVF through long-term (decades) follow-up.

Reproduction is also unique because selection determines who will come into existence. This creates the so-called 'non-identity problem' which has spawned decades of unresolved philosophical debate, sparked by Parfit (1984). Imagine Embryo A has a higher chance of implantation but unknowingly a higher chance of cancer later in life than embryo B. AI selects A. A is born but gets cancer at the age of 30 years. Was A harmed by the decision to select A rather than B? No, a different person (B) would have been otherwise selected. Provided that the disadvantageous trait or genes do not make A's life so bad as to have been not worth living, then A cannot be harmed by selection. On this ground, greater risks can be taken in embryo selection than with interventions on a specific embryo (such as gene editing of A) which do risk harm to a specific individual (Savulescu et al., 2006). Nonetheless, some have argued that parents (and clinicians) still have a moral obligation to select the embryo with the best chance of the best life (Savulescu, 2001; Savulescu and Kahane, 2009, 2016).

Societal impacts of AI for embryo selection

Successful AI models might be deployed at scale, and if such models systematically favor certain traits represented in early morphokinetic profiles, this might impact society. For example, bias to one sex could lead to a skewed population ratio. Similarly, if AI-assisted IVF works better for some races than others, this could have serious societal implications. The scale of these ramifications will likely correlate with rates of IVF use in the future. Since black-box models do not aim to identify specific aspects of an embryo with specific traits, it might make these issues more difficult to detect until it may be too late and there are societal-level impacts. Interpretable AI may allow earlier detection of systematic favoring of certain traits (for instance, if the AI model is known to leverage factors that differ among ethnic groups, e.g., the relation of age to fertility).

Black-box models pose a responsibility gap

The final ethical issue concerns a potential erosion of ethical and legal accountability through the use of opaque AI models. If it is determined that clinicians cannot be held responsible for injuries sustained by the patient due to a reliance on opaque AI models, the responsibility would need to be borne by another agent. In the absence of institutionalized accountability mechanisms that hold other agents, like model developers, responsible, this creates a 'responsibility gap'.

The most straightforward case in which accountability is required would be repeated implantation failure or low success rates due to suboptimal embryo selection processes, and/or injury being sustained by the patient as a result of AI (either to the mother through surgical complications or the child when he/she is born—wrongful life or birth). If AI models used for embryo selection reason in uninterpretable ways, it is unclear how a court might evaluate the doctor's

decision-making, and subsequently, it would be unclear how responsibility for injury would be adjudicated (Price II *et al.*, 2019; Schönberger, 2019).

Legal and regulatory requirements

There is a legal requirement in the European Union's General Data Protection Regulation (GDPR) to provide the patient with 'meaningful information about the logic involved' in automated decisions (Blackmer, 2018). While there is legal debate about interpretation of these regulations (Selbst and Powles, 2017; Wachter *et al.*, 2018), there are increasing calls for automated decisions to be interpretable and explainable to the data subject (Ordish *et al.*, 2020).

The problem of randomized controlled trials of black-box AI algorithms

Within the context of randomized controlled trials (RCTs) of black-box models, it should be noted that they may have accountability implications for poor outcomes in research settings. If an RCT of a black-box fails and the model causes harm to the treatment group, it becomes difficult to ascertain through existing accountability mechanisms who ought to be held responsible.

However, to date, no AI studies for embryo selection using an RCT have been published, though one is registered as a non-inferiority trial (Australian New Zealand Clinical Trials Registry, 2020). It is therefore premature to implement AI-assisted embryo selection in the clinical setting. The lack of RCTs appears to be typical of much of AI in medicine (Nagendran *et al.*, 2020). The problem of lack of evidence before implementation is exacerbated by the IVF industry, which is notorious for aggressively marketing unproven clinical and laboratory 'add-ons' (Wilkinson *et al.*, 2019; Afnan *et al.*, 2020). Furthermore, the introduction of any new tool for embryo selection, without proper assessment and follow-up, raises important ethical issues, as it will determine who will come into existence, and would influence future demographics should the tool be biased toward, say, one sex or ethnicity. The problem is further compounded because clinicians who do not have an adequate understanding of AI will find it difficult to critically navigate the literature, which contains unfamiliar concepts and terminology.

Notwithstanding our concerns about black-box AI models, should such a model be shown (in an RCT) to achieve higher live birth rates than with interpretable models, there is then a good reason to employ them. However, in the absence of such evidence, or if the outcomes are comparable, interpretable ML models are clearly preferable as they are safer than black-box models.

Interpretable ML as the way forward in embryo selection

While we acknowledge that black-box algorithms may be easier and quicker to generate than interpretable algorithms, we strongly advocate for using interpretable ML models to assist embryo selection, as they are safer. As long as one can design the interpretability metric carefully to match the domain, interpretable models tend not to lose accuracy relative to their black-box counterparts (Chen *et al.*, 2019; Rudin, 2019). Interpretable AI may be superior to black-box models in the following ways:

- Confounders are often difficult to detect in evaluation, and they reduce the ability of an algorithm to generalize to populations outside the evaluation population. A clinician may notice them immediately if the model were interpretable.
- An altogether erroneous reasoning process might be easily detected before the tool makes poor selection choices. For instance, after a change in camera setting, an algorithm might suddenly start thinking that the shape of a current embryo looks like an embryo from the training set with a completely different shape. A clinician could potentially catch that problem immediately if they knew the reasoning process of the model.
- Interpretable models would be more robust to deployment across different settings, with different equipment and culture media, if clinicians could modulate their interpretation of an algorithm's recommendation under different conditions. AI has the classic problem of 'domain adaptation'. Once the algorithm shifts to another setting, factors that were helpful previously may not be reliable in the new setting. We cannot assume the AI method is using the information we think it is using, or that we would like it to use. For instance, there are cases in which AI for predicting hospital outcomes looked at the type of X-ray equipment rather than the medical content of the X-ray because the type of equipment actually provided information (sicker patients got portable X-ray images taken more often). Confounding can be very powerful. Interpretable AI is one way to discover such issues. This would make clinics less economically beholden to companies who sell this equipment.
- Physicians and interpretable ML models can create a 'centaur' that leverages both the information in a database (through ML) and a human's system-level way of thinking about problems.
- An interpretable ML model could enhance the shared clinician decision-making process of doctors and patients, instead of replacing it.
- Responsibility for decision-making clearly remains with the clinician. Embryologists monitor embryo morphology on the day of transfer, and morphokinetics during the culture period, and of course ultimately the outcome of pregnancy rates and live births. Should a problem be found, the embryologist will be able to interrogate an interpretable model to troubleshoot the cause of the discrepant findings and make a clinical judgment of not only what the problem is, but how to correct it. With a black-box model, this is not possible. Interpretable models, in which the embryologist's expertise and the ML algorithm are both informative should, in the long run, save time and be more accurate.
- Correlating patient values (chance of disability, sex, single versus double embryo transfer and chance of implantation) with outcomes can be more easily accommodated by interpretable models.

Recommendations

Interpretable AI

Developers should aim to build interpretable ML models where biologically meaningful parameters guide embryo assessment,

reducing the risk of hidden biases in algorithms causing unintended harms to society, permitting better troubleshooting, and better enabling clinicians to counsel their patients on the thinking underlying their treatment.

Randomized controlled trials

Before clinical adoption, the benefits and risks of implementing AI for embryo selection must be studied using the gold-standard for evaluating safety and effectiveness: the RCT.

Regulatory oversight of interpretable AI

Current regulatory approaches attempt to capture medical AI models as a type of medical device; they should further require either that AI model developers not produce black-box models if interpretable models are shown to have similar performance, or that any black-box model must come with the next-best interpretable model considered and trialed. A 'hard' regulatory stance that promotes interpretable models would be safer.

Access to code and data

Data and code used to create ML models should be made publicly accessible. This would enable reproducible research and the advancement of an exciting and important academic field. A high-quality public model would provide a performance baseline for other models. As far as we know, there is not currently a high-quality publicly available dataset for studying embryo selection or implantation. We hope that professional societies and the scientific community make such datasets available, as has occurred in other areas of medicine, such as breast cancer screening, in which large-scale datasets have been made publicly available, funded by the US National Institutes of Health (Buda *et al.*, 2021).

Respect for patient privacy and autonomy

Procedures should be put in place for securing patient privacy when data is shared, such as data anonymization. Currently, the clinician explains the basis on which embryos are selected to the patient, using decision aids like the Gardner scale, which uses biologically meaningful parameters. Ideally, all patients should be told the benefits (e.g., the chance of pregnancy), as well as the limitations (for example being unable to guarantee a healthy baby) and the unknowns (e.g., possible obstetric risks), and how a model arrives at a recommendation (e.g., the number, symmetry, and regularity of the cells). Where possible, patient values should be incorporated. Explaining how a model arrives at a recommendation and incorporating patient values would be difficult if we use black-box models.

Careful and long-term follow-up of babies born after AI-assisted embryo selection

Researchers, professional societies and regulators in a number of countries have recognized the importance of monitoring the health and safety of children born after the introduction of new technologies in assisted reproduction. This applies to the introduction of all novel interventions, such as ICSI and preimplantation genetic testing, and also to AI. While this may be a burden to families, it is necessary following clinical trials, which must inevitably be limited in size, until the technology can be confidently said to be safe and effective.

We summarize our recommendations in Table I.

Table I Summary box of recommendations for use of AI in embryo selection.

- Use of replicable, interpretable ML tools and data
- Well designed and conducted RCTs
- Post implementation surveillance
- Regulatory oversight requiring interpretable AI whenever possible
- Funding for public institutions to transparently develop and evaluate ML models, and open access to code used in models
- Procedures for maintaining security of patient/embryo data while permitting ethical data sharing
- Fully informed consent to use AI
- Inclusion of patient values into AI programs where possible
- Training for clinicians to understand AI models and explain them to patients

AI, artificial intelligence; ML, machine learning; RCT, randomized controlled trial.

Conclusion

We do not aim to demonize AI. Quite the opposite, we enthusiastically acknowledge that it has the potential to radically enhance IVF. AI in IVF has the potential to help couples have children earlier in their treatment and at a lower cost. The point is that AI algorithms are portrayed as applications which are so clever that they can detect order from background noise that is too subtle for the human to detect. On the flip side is the risk that the algorithm will find order when none exists or is there by association. These are two sides of the same coin. Researchers, companies and clinics must ensure that the technology they promote or adopt brings real, measurable benefits to patients and, most importantly, does not expose them to unreasonable risks. We have highlighted limitations of current ML models, we drew specific attention to the ethical concerns associated with AI in IVF, and suggested changes to design and evaluation. We have argued that there should be interpretable ML models that clinicians can understand, troubleshoot and explain to their patients, rigorously evaluated with RCTs. Black-box AI risks a new machine paternalism.

Supplementary data

Supplementary data are available at *Human Reproduction Open* online.

Data availability

No new data were generated or analyzed in support of this research.

Authors' roles

All authors contributed equally to this paper.

Funding

This research received no grant from any funding agency.

Conflict of interest

The authors declare no conflicts of interest.

References

Afnan MAM, Khan KS, Mol BW. Generating translatable evidence to improve patient care: the contribution of human factors. *Reprod Biomed Online* 2020;**41**:353–356.

Afnan MAM, Rudin C, Conitzer V, Savulescu J, Mishra A, Liu Y, Afnan M. Ethical implementation of artificial intelligence to select embryos in *in vitro* fertilization. In: *AIES 2021—Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics and Society*. 2021, 316–326.

Agar N. *Liberal Eugenics: In Defense of Human Enhancement*. Hoboken: Blackwell Publishing, 2004.

Australian New Zealand Clinical Trials Registry. *Evaluation of the Role of an Artificial Intelligence System (iDA) in Embryo Selection*. 2020. https://www.anzctr.org.au/Trial/Registration/TrialReview.aspx?id=379161&isReview=true (3 November 2021, date last accessed).

Bjerring JC, Busch J. Artificial intelligence and patient-centered decision-making. *Philos Technol* 2020;**34**:349–371.

Blackmer WS. EU general data protection regulation. *OJEU* 2018;**2014**:45–62.

Bronet F, Nogales M-C, Martinez E, Ariza M, Rubio C, Garcia-Velasco J-A, Meseguer M. Is there a relationship between time-lapse parameters and embryo sex? *Fertil Steril* 2015;**103**:396–401.e2.

Buda M, Saha A, Walsh R, Ghate S, Li N, Swiecicki A, Lo JY, Mazurowski MA. A data set and deep learning algorithm for the detection of masses and architectural distortions in digital breast tomosynthesis images. *JAMA Netw Open* 2021;**4**:e2119100.

Centers for Disease Control and Prevention, US Department of Health and Human Services. *2018 Assisted Reproductive Technology Fertility Clinic Success Rates Report*. 2018. http://www.cdc.gov/art/reports (3 November 2021, date last accessed).

Charles C, Gafni A, Whelan T. Shared decision-making in the medical encounter: what does it mean? (Or it takes at least two to tango). *Soc Sci Med* 1997;**44**:681–692.

Chavez-Badiola A, Flores-Saiffe Farias A, Mendizabal-Ruiz G, Garcia-Sanchez R, Drakeley AJ, Garcia-Sandoval JP. Predicting pregnancy test results after embryo transfer by image feature extraction and analysis using machine learning. *Sci Rep* 2020;**10**:4394.

Chen C, Li O, Tao D, Barnett A, Rudin C, Su JK. This looks like that: deep learning for interpretable image recognition. In: *Proceedings of the 33rd Conference on Neural Information Processing Systems*, Vol. 32. 2019, 8930–8941.

Curchoe CL, Flores-Saiffe Farias A, Mendizabal-Ruiz G, Chavez-Badiola A. Evaluating predictive models in reproductive medicine. *Fertil Steril* 2020;**114**:921–926.

Elwyn G, Edwards A, Wensing M, Hood K, Atwell C, Grol R. Shared decision making: developing the OPTION scale for measuring patient involvement. *Qual Saf Health Care* 2003;**12**:93–99.

Fernandez EI, Ferreira AS, Cecilio MHM, Cheles DS, Souza R. D, Nogueira MFG, Rocha JC. Artificial intelligence in the IVF laboratory: overview through the application of different types of algorithms for the classification of reproductive data. *J Assist Reprod Genet* 2020;**37**:2359–2376.

Gardner DK, Lane M, Stevens J, Schlenker T, Schoolcraft WB. Blastocyst score affects implantation and pregnancy outcome: towards a single blastocyst transfer. *Fertil Steril* 2000;**73**:1155–1158.

Gardner DK, Meseguer M, Rubio C, Treff NR. Diagnosis of human preimplantation embryo viability. *Hum Reprod Update* 2015;**21**:727–747.

Garvelink MM, Boland L, Klein K, Nguyen DV, Menear M, Bekker HL, Eden KB, LeBlanc A, O'Connor AM, Stacey D *et al*. Decisional conflict scale use over 20 years: the anniversary review. *Med Decis Making* 2019;**39**:301–314.

HFEA. *Fertility Treatment 2018: Trends and Figures Quality and Methodology Report* | Human Fertilisation and Embryology Authority. 2020. https://www.hfea.gov.uk/about-us/publications/research-and-data/fertility-treatment-2018-trends-and-figures/fertility-treatment-2018-quality-and-methodology-report/ (3 November 2021, date last accessed).

Huang B, Ren X, Zhu L, Wu L, Tan H, Guo N, Wei Y, Hu J, Liu Q, Chen W *et al*. Is differences in embryo morphokinetic development significantly associated with human embryo sex? *Biol Reprod* 2019;**100**:618–623.

Kan-Tor Y, Ben-Meir A, Buxboim A. Can deep learning automatically predict fetal heart pregnancy with almost perfect accuracy? *Hum Reprod* 2020;**35**:1473–1473.

Kemper JM, Wang R, Rolnik DL, Mol BW. Preimplantation genetic testing for aneuploidy: are we examining the correct outcomes? *Hum Reprod* 2020;**35**:2408–2412.

Khosravi P, Kazemi E, Zhan Q, Malmsten JE, Toschi M, Zisimopoulos P, Sigaras A, Lavery S, Cooper LAD, Hickman C *et al*. Deep learning enables robust assessment and selection of human blastocysts after *in vitro* fertilization. *Npj Digit Med* 2019;**2**:1–9.

Kramer MF, Schaich Borg J, Conitzer V, Sinnott-Armstrong W. When do people want AI to make decisions? In: *AIES 2018—Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics and Society*. 2018, 204–209.

Leahy BD, Jang W-D, Yang HY, Struyven R, Wei D, Sun Z, Lee KR, Royston C, Cam L, Kalma Y *et al*. Automated measurements of key morphological features of human embryos for IVF. *Med Image Comput Comput Assist Interv* 2020;**12265**:25–35.

Légaré F, Kearing S, Clay K, Gagnon S, D'Amours D, Rousseau M, O'Connor A. Are you SURE? Assessing patient decisional conflict with a 4-item screening test. *Can Fam Physician* 2010;**56**:e308–e314.

Liu Y, Chapple V, Feenan K, Roberts P, Matson P. Time-lapse deselection model for human day 3 *in vitro* fertilization embryos: the combination of qualitative and quantitative measures of embryo growth. *Fertil Steril* 2016;**105**:656–662.e1.

Liu Y, Feenan K, Chapple V, Matson P. Assessing efficacy of day 3 embryo time-lapse algorithms retrospectively: impacts of dataset type and confounding factors. *Hum Fertil (Camb)* 2019;**22**:182–190.

McDougall RJ. Computer knows best? The need for value-flexibility in medical AI. *J Med Ethics* 2019;**45**:156–160.

Milewski R, Kuczyńska A, Stankiewicz B, Kuczyński W. How much information about embryo implantation potential is included in morphokinetic data? A prediction model based on artificial neural networks and principal component analysis. *Adv Med Sci* 2017;**62**: 202–206.

Mishra A, Savulescu J, Giubilini A. Ethics of medical AI. The Oxford Handbook of Ethical Theory. In Press.

Morales DA, Bengoetxea E, Larranaga P, Garcia M, Franco Y, Fresnada M, Merino M. Bayesian classification for the selection of *in vitro* human embryos using morphological and clinical data. *Comput Methods Programs Biomed* 2008;**90**:104–116.

Nagendran M, Chen Y, Lovejoy CA, Gordon AC, Komorowski M, Harvey H, Topol EJ, Ioannidis JPA, Collins GS, Maruthappu M. Artificial intelligence versus clinicians: systematic review of design, reporting standards, and claims of deep learning studies. *BMJ* 2020; **368**:m689.

O'Connor M. Algorithm's 'unexpected' weakness raises larger concerns about AI's potential in broader populations. *Health Imaging*. 2021. https://www.healthimaging.com/topics/artificial-intelligence/weakness-ai-broader-patient-populations?utm_source=newsletter&utm_medium=rb_news (3 November 2021, date last accessed).

Ordish J, Brigden T, Hall A. Black box medicine and transparency | PHG Foundation. 2020, 34. https://www.phgfoundation.org/research/black-box-medicine-and-transparency (3 November 2021, date last accessed).

Parasuraman R, Manzey DH. Complacency and bias in human use of automation: an attentional integration. *Hum Factors* 2010;**52**:381–410.

Parfit D. *Reasons and Persons*. Oxford: Oxford University Press, 1984.

Price W II, Gerke S, Cohen IG. Potential liability for physicians using artificial intelligence. *JAMA* 2019;**322**:1765–1766.

Raef B, Maleki M, Ferdousi R. Computational prediction of implantation outcome after embryo transfer. *Health Inform J* 2020;**26**:1810–1826.

Rosenwaks Z. Artificial intelligence in reproductive medicine: a fleeting concept or the wave of the future? *Fertil Steril* 2020;**114**:905–907.

Rudin C, Chen C, Chen Z, Huang H, Semenova L, Zhong C. Interpretable machine learning: fundamental principles and 10 grand challenges. 2021;1–74.http://arxiv.org/abs/2103.11251.

Rudin C, Radin J. Why are we using black box models in AI when we don't need to? A lesson from an explainable AI competition. *Harvard Data Sci Rev* 2019;**1**:1–9.

Rudin C. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nat Mach Intell* 2019;**1**:206–215.

Savulescu J, Hemsley M, Newson A, Foddy B. Behavioural genetics: why eugenic selection is preferable to enhancement. *J Appl Philos* 2006;**23**:157–171.

Savulescu J, Kahane G. The moral obligation to create children with the best chance of the best life. *Bioethics* 2009;**23**:274–290.

Savulescu J, Kahane G. Understanding procreative beneficence. In Francis L, editor. *The Oxford Handbook of Reproductive Ethics*. 2016; Oxford, UK: Oxford University Press.

Savulescu J. Procreative beneficence: why we should select the best children. *Bioethics* 2001;**15**:413–426.

Schönberger D. Artificial intelligence in healthcare: a critical analysis of the legal and ethical implications. *Int J Law Inf Technol* 2019;**27**: 171–203.

Selbst AD, Powles J. Meaningful information and the right to explanation. *Int Data Priv Law* 2017;**7**:1–20.

Tarín JJ, García-Pérez MA, Hermenegildo C, Cano A. Changes in sex ratio from fertilization to birth in assisted-reproductive-treatment cycles. *Reprod Biol Endocrinol* 2014;**12**:56.

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**:1011–1018.

Voter AF, Meram E, Garrett JW, Yu J-PJ. Diagnostic accuracy and failure mode analysis of a deep learning algorithm for the detection of intracranial hemorrhage. *J Am Coll Radiol* 2021;**18**:1143–1152.

Wachter S, Mittelstadt B, Russell C. Counterfactual explanations without opening the black box: automated decisions and the GDPR. *New Media Soc* 2018;**20**:973–989.

Wang A, Kort J, Behr B, Westphal LM. Euploidy in relation to blastocyst sex and morphology. *J Assist Reprod Genet* 2018;**35**: 1565–1572.

Wilkinson J, Malpas P, Hammarberg K, Mahoney Tsigdinos P, Lensen S, Jackson E, Harper J, Mol BW. Do à la carte menus serve infertility patients? The ethics and regulation of *in vitro* fertility add-ons. *Fertil Steril* 2019;**112**:973–977.

Zmuidinaite R, Sharara FI, Iles RK. Current advancements in noninvasive profiling of the embryo culture media secretome. *Int J Mol Sci* 2021;**22**:1–13.